

For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex LIBRIS
UNIVERSITATIS
ALBERTAEENSIS



THE UNIVERSITY OF ALBERTA

AUTOMATIC DOCUMENT CLASSIFICATION SYSTEMS

by



SHIGEKO AKIYAMA

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

FALL, 1972

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "AUTOMATIC DOCUMENT CLASSIFICATION SYSTEMS" submitted by SHIGEKO AKIYAMA in partial fulfilment of the requirements for the degree of Master of Science.

Date...*May 26th*...1972

ABSTRACT

The present thesis examines a technique for automatically classifying documents according to their subject categories. Experiments are described for a data base of 1572 titles of papers published by the Journal of Acoustical Society of America in 1966, 1967, 1968, and 1961.

The feasibility of using latent class analysis for the document classification is tested by two experiments. The technique proposed by F. B. Baker and W. K. Winters is found to be unsuitable for practical application to document classification, because the matrices required by the theory to be positive definite are in fact found to be non-positive definite. Another attempt to solve the accounting equations that describe the latent class structure is based on the optimization technique. This method requires an enormous amount of computation time and still does not determine meaningful classes. It is concluded that latent class analysis is not a useful technique for solution of the problem of document classification.

The classification method based on attribute analysis proposed by M. E. Maron is applied to the classification of the acoustical literature. With use of a proposed procedure for choice of keywords from document titles the results appear to be very satisfactory. In particular, Maron's assumption that keywords of a document occur in a statistically independent manner does not appear to reduce the effectiveness of the classification.

A modified application of attribute analysis to document classification is proposed through maximization of correct classifications

of base documents using not more than two keywords in the computation of joint word occurrences, but without use of approximate estimates. The results are slightly superior to those of Maron's method.

Acknowledgment

The author wishes to extend her appreciation to

- Professor H. S. Heaps for his supervision of this thesis,
- Mrs. Evelyn Buchanan for her patience in typing this thesis,
- The University of Alberta for teaching assistantship,
- Her parents and husband for their constant encouragement.

TABLE OF CONTENTS

	<u>PAGE</u>
CHAPTER I INTRODUCTION	1
1.1 General	1
1.2 Statistical Analysis	4
1.3 Latent Class Analysis	5
1.4 Statement of the Approach of Subsequent Chapters	6
CHAPTER II LATENT CLASS ANALYSIS	9
2.1 General	9
2.2 Latent Class Structure	10
2.3 Numerical Solution of Winters	13
2.4 Summary	19
CHAPTER III APPLICATION OF LATENT CLASS ANALYSIS	20
3.1 Application of Winters' Method Using Experimental Data	20
3.2 An Attempt to Use Latent Class Analysis	23
3.2.1 General	23
3.2.2 Numerical Solution	23
3.2.3 Application of Proposed Method . .	26
3.2.4 Discussion	29
3.3 Conclusions Regarding the Limitations, or Unsuitability, of Latent Class Determination	32
3.3.1 Winters' Method	32
3.3.2 Minimizing Method	34
3.3.3 Summary	34

	<u>PAGE</u>
CHAPTER IV	ATTRIBUTE ANALYSIS 36
	4.1 Classification by Attribute Number . . 36
	4.2 Selection of Data 37
	4.3 Selection of Categories 38
	4.4 Selection of Keywords 38
	4.5 Experimental Results 40
	4.6 Summary 41
CHAPTER V	ACOUSTICS DATA BASE AND SELECTION OF
	KEYWORDS 44
	5.1 Selection of Data 44
	5.2 Selection of Categories 44
	5.3 Selection of Keywords 46
	5.4 Statistics on Data 49
CHAPTER VI	APPLICATION OF MARON'S ATTRIBUTE ANALYSIS TO
	ACOUSTICS DATA BASE 53
	6.1 Experimental Results on Acoustic Data . 53
	6.2 Discussion of Results 55
	6.3 Possible Improvements in Procedure . . 56
CHAPTER VII	MODIFIED ATTRIBUTE ANALYSIS 58
	7.1 Maximization of Correct Document
	Classifications 58
	7.1.1 Classification System 58
	7.1.2 Experimental Results 62
	7.1.3 Suggestions and Discussion 66
	7.2 Modification of Maron's Method Using
	Keyword Association 67

	<u>PAGE</u>
7.2.1 Classification System Based on	
Keyword Association	67
7.2.2 Experimental Results	70
7.2.3 Discussion	70
7.2.4 Suggestion	72
CHAPTER VIII CONCLUSIONS	75
REFERENCES	78
APPENDIX A List of Keywords Chosen as Described in	
Chapter V	82
APPENDIX B Similarity of Successive Years of Data Base .	84
APPENDIX C Stiles' Measure of Association Factor and	
Choice of Keywords	86
APPENDIX D First and Second Rank Classification Using	
Modified Attribute Analysis	89

LIST OF TABLES

	<u>PAGE</u>
Table 3.1 A Sample Solution of the Accounting Equations	27
Table 3.2 Approximates Given by Minimizing Method . . .	30
Table 3.3 Comparison of Postulated g^{ℓ} and x_i^{ℓ} with Computed Values (in parentheses)	31
Table 4.1 Summary of Maron's Results	42
Table 5.1 Word Frequency Table Used for Keyword Selection	48
Table 5.2 Distribution of Number of Titles in Group 1 (1966) over 14 Categories	50
Table 5.3 Distribution of Number of Titles in Group 2 (1967) over 14 Categories	50
Table 5.4 Number of Keywords in Titles	52
Table 6.1 Experimental Results of Maron's Method Applied to Acoustics Data	54
Table 7.1 An Example of Single Keyword Table (consisted in the instance of only two keywords and three categories)	61
Table 7.2 Experimental Results of Maximization Method .	62
Table 7.3 Comparison of Maximization Method with Maron's Method	65
Table 7.4 Experimental Results of Modified Method and Their Comparison with Those of Maron's Method	71

CHAPTER I

INTRODUCTION

1.1 General.

In recent years a number of investigations and experiments have been undertaken in various aspects of automatic documentation. They have dealt with the structure, analysis, organization, storage, search, and retrieval of information. As a result, the conceptual analysis of documents has become a basic consideration in document handling.

In conventional library systems trained people analyze the subject matter of documents and either assign index words to them or else classify them in accordance with existing hierarchical classification schedules. At the present time the rate of growth of documentary data is sufficiently high that many libraries face serious problems concerning the size of storage media, the method of file organization, and the education of skillful librarians. As a result of increases in the quantity of information there are strong demands for the creation of services to supply needed information that is directly, or indirectly, related to the interests of particular researchers. However, it is very time consuming to handle mass information manually because many research subjects are not limited to narrow fields; but tend to spread over other related fields.

In many automatic documentation systems the storage of information is not the main problem. It may be solved by provision of sufficient hardware devices such as magnetic tapes, discs, drums, magnetic cards, and microfilm, and so forth. Much manual work may be eliminated by use of mechanization. Furthermore, the use of computers allows more

sophisticated document processing such as automatic retrieval, abstracting, indexing, and classification. However, even with use of automation there still remain serious problems in the analysis and the identification of content.

In the early 1960's G. Salton and his group at Harvard University designed the system known as SMART, Salton's Magical Automatic Retrieval Technique (17, 18). It is a fully mechanized information system and is in operation at Harvard and Cornell Universities. The outstanding feature of the SMART system is that it may use several hundred different forms of content analysis in order to determine the correct words that should be used to represent and search documents. The techniques include use of a thesaurus, statistical word associations, syntactic analysis, statistical phrase recognition, and hierarchical arrangement of concepts. Implementation of the SMART system has helped to prove the practical feasibility of automatic information processing.

According to Richardson's definition, (16) "classification" is the putting together of like things. Every entity, nature, idea, and art may be analyzed and classified in accordance with appropriate classification schedules. The present thesis, however, concentrates on the classification of scientific documents that are described by natural language such as used in titles, abstracts, keywords, and subject headings.

There exist general classification schemes such as the Universal Decimal Classification (UDC) (22), the Dewey Decimal Classification (DC) (4), the Library of Congress Classification (LC) (9), and the Colon Classification (CC) (14). They are not satisfactory enough for classification of highly specialized subjects because they do not sufficiently represent the details of a complex subject, and they do not provide

sufficient flexibility in classification of documents that relate to several fields. In order to overcome these disadvantages, "Faceted Classification" was developed by Vickery (23), and "Analytico-Synthetic Classification" by Ranganathan (15). For these classifications the main facets in each subject field must be generated. For example, in the subject field of Food Technology there may be four facets, Products, Parts, Materials, and Operations, and these main facets may be further divided into sub-facets and sub-sub-facets, and so forth. Obviously these techniques make it possible to analyze the document concepts in greater depth.

When adapted to automated systems the existing general classification schemes referred to above require considerable help from human beings since the conceptual analysis of documents is performed manually. One of the aims of research in the field of document classification is to clearly understand the relationships between document content and assigned subject categories. With such an understanding it is hoped that subject categories may be assigned automatically by computer examination of the document content.

The extent to which subject categories may be chosen by automatic examination of document content is the subject of the present thesis. Attention is confined to examination of titles only. Comparison is made with the results of manual classification based solely on examination of titles. Accordingly, the aim of the present investigation is to compare and evaluate several methods of automatic classification, to modify them if necessary, and to compare their effectiveness with that of manual classification.

measure of the degree of correlation of words in terms of their frequencies of occurrence, and he attempted to formulate the means to calculate it automatically in terms of the association factor.

1.3 Latent Class Analysis.

Latent class analysis was first introduced by Lazarsfeld (8) for application in the field of social psychology in order to analyze a set of questionnaires to assess the attitude of army personnel in terms of various factors. The analysis is based on a mathematical model based on the assumption that a set of data described by statistics may be divided into small sets such that in each group the probabilities of different word incidences are statistically independent.

In that statistical independence of incidences it is assumed within any group both latent class analysis and attribute analysis are essentially the same. However, there is considerable difference in the procedure used to derive the estimates of the necessary probabilities. In attribute analysis the probabilities which are used to predict the attribute of a whole are derived from a pre-existing relatively small amount of data which has already been classified. On the other hand, in latent class analysis, the probabilities are generated directly from the attributes. The advantage of latent class analysis is that the automatic classification groups may be derived from the automatic generation of the latent class structure, whereas when based on attribute analysis, the groups depend on a previously chosen set of categories.

In 1954, T. W. Anderson (1) proposed a method for the numerical solution of certain equations that involve probabilities and which arise in construction of the latent class model. The Anderson technique was

developed to overcome the inherent difficulty of the method suggested earlier by B. Green (6), in which the values of the elements of certain required matrices cannot be defined precisely, and hence must be approximated. Anderson formed square matrices of elements that represent correlation probabilities of keywords, and he applied eigenvalue techniques. However, he did not note that asymmetric matrices do not necessarily have real eigenvalues.

In 1962, F. B. Baker (2, 3) first realized that the latent class structure may be directly applied to the field of document classification and, in fact, could be used to provide the necessary mathematical foundation for a method of automatic classification.

The difficulties that arise through introduction of asymmetric matrices may be overcome by use of the latent class formulation proposed by Winters (24). It is a modification of Anderson's technique, and leads to generation of symmetric matrices and hence real eigenvalues. The elements of Winters' matrices represent probabilities of occurrences of single keywords, double keywords, and triple keywords. Use of combinations of more keywords may construct a firmer latent class model, but the probabilities of such combinations become small or zero, and may be neglected in practice in the construction of latent classes.

In application of the method of Winters, eigenvalues are required to describe probabilities. Winters did not discuss the conditions required to ensure that the eigenvalues lie between 0 and 1; yet this condition is essential if the eigenvalues are to represent probabilities.

1.4 Statement of the Approach of Subsequent Chapters.

The purpose of the present thesis is to critically examine and, if

necessary, develop the methods of statistical analysis for automatic classification in terms of association of keywords and subject categories.

Chapter II contains a discussion of latent class analysis, with emphasis on consideration of the practicality of the method of Winters in so far as the required numerical computations are concerned. An experimental attempt to apply latent class analysis to an existing document data base is described in Chapter III. It is demonstrated that, contrary to the hopes of Baker, the method of latent class analysis is not suitable for automatic determination of document categories. Chapter IV contains a discussion of attribute analysis and the experimental results obtained by Maron.

The experimental results described in the present thesis were obtained by use of a data base that contains references to journal articles in the field of acoustics. The data base and its subject categories are described in Chapter V.

Application of Maron's method of attribute analysis is made in Chapter VI. Although the method is not new it is believed that the results are of value in providing assessment of Maron's method, since the data base is much larger than that used by Maron, and therefore it provides a more realistic example of a document data base. Furthermore, the categories used by Maron were the result of his modification of an existing classification scheme, whereas the categories used in the present experiment are ones that have been in use since 1961. It is therefore believed that the experimental results provide a useful measure of the effectiveness of Maron's method in comparison with a well established and accepted method of manual assignment of categories.

The classification obtained in Chapter VI is based on use of keywords

chosen from document titles whereas the results of Maron were based on use of keywords selected from document abstracts. The results of Chapter VI indicate that the method of choice of keywords from titles leads to automatic classification that is as good as that obtained by Maron when using keywords chosen from abstracts.

In Chapter VII there are introduced some modifications of attribute analysis. The results are compared with those of Chapter VI.

CHAPTER II

LATENT CLASS ANALYSIS

2.1 General.

In application of latent class analysis to automatic document classification systems it is supposed that, within an entire corpus of documents, there exists a set of non-intersecting classes in which the occurrence of each keyword in a document is statistically independent of the occurrences of other keywords. The latent class analysis then proceeds through use of probabilities that describe associations between latent classes and certain combinations of keywords. The associations are formulated in the form of probabilities that a document with a particular combination of keywords belongs to any of latent classes.

F. B. Baker (2) first attempted to apply techniques employed by Lazarsfeld (8). He proposed the mathematical model of latent class analysis, and suggested how to use it.

W. K. Winters (24) modified Baker's latent class structure and discussed the numerical procedures required.

Use of a large number of keywords allows the numerical methods to determine close approximations to the latent classes, but because of the complexity of the computations the number of considered combinations of keywords must be limited. Furthermore, the most difficult problem involved in the construction of latent classes is determination of the number of classes to be sought. It seems that there is no firm theory to determine

this number. Baker proposed the inequality $(N + 1) / 2 > L$ between the number of keywords and the number of classes, where N denotes the number of keywords and L denotes the number of classes. However, the manner in which he obtained this inequality is not explained. In order to make the numerical solution feasible in practice, Winters assumed that the number of latent classes is equal to the number of existing keywords, that is, $L = N$. Clearly Winters' assumption is in disagreement with the Baker inequality, and there is a need for further investigation before either relation between N and L may be used with any degree of confidence.

2.2 Latent Class Structure

The latent class analysis used the following probabilities for keyword occurrences in the entire set of documents:

$$\begin{aligned} p_i &= \text{probability that a document contains the keyword } K_i, \\ p_{ij} &= \text{probability that a document contains both keywords } K_i \text{ and } K_j, \\ p_{ijk} &= \text{probability that a document contains three keywords } K_i, \\ &\quad K_j, \text{ and } K_k. \end{aligned}$$

For a document that belongs to the latent class C_ℓ , the probabilities h_i^ℓ , h_{ij}^ℓ , and h_{ijk}^ℓ are defined as follows:

$$\begin{aligned} h_i^\ell &= \text{probability that the document contains the keyword } K_i, \\ h_{ij}^\ell &= \text{probability that the document contains both keywords} \\ &\quad K_i \text{ and } K_j, \\ h_{ijk}^\ell &= \text{probability that the document contains three keywords} \\ &\quad K_i, K_j, \text{ and } K_k. \end{aligned}$$

For an arbitrary document chosen from the entire data the probability

g^ℓ is defined as follows:

g^ℓ = probability that the document belongs to the latent class C_ℓ .

With N keywords and L latent classes, there are 2^N p's, L g's and $L \times 2^N$ h's. The relationships between the p's, g's, and h's are expressed by the following equations, known as the accounting equations:

$$p_i = \sum_{\ell=1}^L g^\ell h_i^\ell \quad (2.1)$$

$$p_{ij} = \sum_{\ell=1}^L g^\ell h_{ij}^\ell \quad (\text{for } i \neq j) \quad (2.2)$$

$$p_{ijk} = \sum_{\ell=1}^L g^\ell h_{ijk}^\ell \quad (\text{for } i \neq j, i \neq k, \text{ and } j \neq k) \quad (2.3)$$

etc.

There is one more equation, namely

$$1 = \sum_{\ell=1}^L g^\ell \quad (2.4)$$

which expresses the fact that a document belongs to one, and only one, latent class.

Basically the problem of latent class analysis is to find the solution for the g's and h's to satisfy the accounting equations (2.1), (2.2), (2.3), etc., and (2.4).

The defining property of a latent class is that, for all documents within it, the keywords occur in a statistically independent manner so that

$$h_{ij}^\ell = h_i^\ell h_j^\ell \quad h_{ijk}^\ell = h_i^\ell h_j^\ell h_k^\ell \quad \text{etc.}$$

The accounting equations may then be rewritten in the form

$$p_i = \sum_{\ell} g^{\ell} h_i^{\ell} \quad (2.5)$$

$$p_{ij} = \sum_{\ell} g^{\ell} h_i^{\ell} h_j^{\ell} \quad (i \neq j) \quad (2.6)$$

$$p_{ijk} = \sum_{\ell=1}^L g^{\ell} h_i^{\ell} h_j^{\ell} h_k^{\ell} \quad (i \neq j, i \neq k, \text{ and } j \neq k) \quad (2.7)$$

etc..

Once the unknown g 's and h 's have been estimated, the degree of association between a document that contains a particular combination of keywords, say K_1, K_2, \dots, K_M , and the latent class C_{ℓ} may be defined as the probability

$$p^{\ell} = \frac{D g^{\ell} h_{1,2 \dots M}^{\ell}}{\sum_{K=1}^L D g^K h_{1,2 \dots M}^K} \quad (2.8)$$

where D is the total number of documents. Then p^{ℓ} is the probability that a document indexed by keywords K_1, K_2, \dots, K_M belongs to the latent class C_{ℓ} .

By the independence assumption, $h_{1,2 \dots M}^{\ell} = h_1^{\ell} h_2^{\ell} \dots h_M^{\ell}$, and hence

$$p^{\ell} = \frac{g^{\ell} h_1^{\ell} h_2^{\ell} \dots h_M^{\ell} (1-h_{M+1}^{\ell}) \dots (1-h_N^{\ell})}{\sum_{K=1}^L g^K h_1^K h_2^K \dots h_M^K (1-h_{M+1}^K) \dots (1-h_N^K)} \quad (2.9)$$

which is computable in terms of the g 's and h 's. The latent class for which this probability assumes its maximum value is the class that is assigned to the given document. This probability is called the "ordering ratio".

2.3 Numerical Solution of Winters.

Under the assumption that the number of keywords equals the number of latent classes, Winters used a modification of T. W. Anderson's technique to propose one possible solution in matrix notation.

Defining five $N \times N$ (or $L \times L$) square matrices as follows:

$$P = \begin{bmatrix} p_N & p_{1,N} & p_{2,N} \cdots & p_{N-1,N} \\ p_{1,N} & p_{1,1,N} & p_{1,2,N} \cdots & p_{1,N-1,N} \\ p_{2,N} & p_{2,1,N} & p_{2,2,N} \cdots & p_{2,N-1,N} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ p_{N-1,N} & p_{N-1,1,N} & p_{N-1,2,N} \cdots & p_{N-1,N-1,N} \end{bmatrix} \quad (2.10)$$

$$\hat{P} = \begin{bmatrix} 1 & p_1 & p_2 \cdots & p_{N-1} \\ p_1 & p_{1,1} & p_{1,2} \cdots & p_{1,N-1} \\ p_2 & p_{2,1} & p_{2,2} \cdots & p_{2,N-1} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ p_{N-1} & p_{N-1,1} & p_{N-1,2} \cdots & p_{N-1,N-1} \end{bmatrix} \quad (2.11)$$

$$H = \begin{bmatrix} 1 & h_1^1 & h_2^1 & \dots & h_{N-1}^1 \\ 1 & h_1^2 & h_2^2 & \dots & h_{N-1}^2 \\ 1 & h_1^3 & h_2^3 & \dots & h_{N-1}^3 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & h_1^N & h_2^N & \dots & h_{N-1}^N \end{bmatrix} \quad (2.12)$$

$$\hat{H} = \begin{bmatrix} h_N^1 & 0 & 0 & \dots & 0 \\ 0 & h_N^2 & 0 & \dots & 0 \\ 0 & 0 & h_N^3 & \dots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \dots & h_N^N \end{bmatrix} \quad (2.13)$$

and

$$G = \begin{bmatrix} g^1 & 0 & 0 & \dots & 0 \\ 0 & g^2 & 0 & \dots & 0 \\ 0 & 0 & g^3 & \dots & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & g^N \end{bmatrix} \quad (2.14)$$

where P and \hat{P} are symmetric, and \hat{H} and G are diagonal matrices, the above accounting equations may be rewritten in the form

$$P = H'G\hat{H}H \quad (2.15)$$

and

$$\hat{P} = H'GH \quad (2.16)$$

where H' denotes the transpose of H . However, these matrix forms (2.15) and (2.16) represent combinations of up to only three keywords in the accounting equations. Because the occurrence of any given set of more than three keywords in a document may be relatively rare, then most of the neglected probabilities are zero or very small, so that neglecting the combinations of more than three keywords should not have any serious effect on the results.

Winters made the assumption that the matrices H , G , and \hat{H} are non-singular. This assumption implies that all the diagonal g 's and h 's must be non-zero. Using this assumption he proved that P and \hat{P} are positive definite so that all eigenvalues of P and \hat{P} are positive.

In order to solve the system described by (2.15) and (2.16) consider

the following generalized eigenvalue problem:

$$P\underline{x} = \lambda \hat{P}\underline{x} \quad (2.17)$$

where \underline{x} is an eigenvector associated with the eigenvalue λ . Defining a matrix T which satisfies the condition of

$$T' \hat{P} T = I \quad (2.18)$$

then the matrix $T'PT$ has eigenvalues equal to the solutions λ of equation (2.17).

The fact can be proved as follows:

Pre-multiply the equation (2.17) by T' to obtain

$$T'P\underline{x} = \lambda T' \hat{P}\underline{x} \quad (2.19)$$

Let $\underline{x} = T\underline{y}$. Substituting it in formula (2.19), we get

$$T'PT\underline{y} = \lambda T' \hat{P} T\underline{y} \quad (2.20)$$

Since $T' \hat{P} T = I$, the equation (2.20) becomes

$$T'PT\underline{y} = \lambda \underline{y} \quad (2.21)$$

and hence λ is an eigenvalue of the matrix $T'PT$.

The purpose of this numerical technique is to derive the unknown matrices H , \hat{H} , and G defined in (2.12), (2.13), and (2.14) respectively.

It may be shown that the diagonal matrix \hat{H} can be obtained by solving the characteristic equation

$$|T'PT - \lambda I| = 0 \quad (2.22)$$

The proof is as follows:

$$\begin{aligned}
0 &= |T'PT - \lambda I| \\
&= |T'PT - \lambda T'\hat{P}T| \\
&= |T'H'G\hat{H}HT - \lambda T'H'GHT| \\
&= |T'| |H'| |G| |\hat{H} - \lambda I| |H| |T|
\end{aligned} \tag{2.23}$$

where H , G , and T are assumed to be non-singular. Therefore,

$$0 = |\hat{H} - \lambda I| \tag{2.24}$$

Thus, since \hat{H} is a diagonal matrix its elements h_N^λ are equal to the eigenvalues λ of equation (2.22).

We note that all eigenvectors \underline{x} 's which are column vectors may be arranged to form a square matrix X . Assume that there exists a diagonal matrix D which satisfies the relation

$$\hat{P}X = H'GD \tag{2.25}$$

The matrices G and D are diagonal so that GD is also a diagonal matrix, and furthermore, the first row of H' consists of all ones. Therefore the diagonal elements of GD must be equal to the elements on the first row of $\hat{P}X$. By post multiplying the formula (2.25) by $(GD)^{-1}$ we may obtain in the form

$$\hat{P}X(GD)^{-1} = H' \tag{2.26}$$

Substituting $\hat{P} = H'GH$, and eliminating H' , the formula (2.26) can be rewritten as follows:

$$GHX(GD)^{-1} = I \tag{2.27}$$

so that finally G may be expressed in the form

$$HX(GD)^{-1} = G^{-1} \tag{2.28}$$

It should be noted that $(GD)^{-1}$ is easily computed by taking the reciprocal of each diagonal element of GD . Similarly, since G is a diagonal

matrix, it may be obtained in a trivial manner from G^{-1} .

The eigenvalue problem defined in (2.22) involves a symmetric matrix, and hence is suitable for attempted solution by either Jacobi's method, Givens' method, or Householder's method.

It remains to determine T which is defined in the formula (2.18). We use an important theorem relative to the eigenvalue problem of symmetric matrices, namely that if a matrix A is symmetric then there exists an orthogonal matrix Q such that

$$Q'AQ = D \quad (2.29)$$

where D is a diagonal matrix whose diagonal elements are the eigenvalues of A . In the formula (2.18) the matrix \hat{P} is indeed symmetric so that by applying a suitable method, the eigenvalues as diagonal elements of D , and the eigenvectors as columns of Q , may be determined to satisfy the equation

$$Q'\hat{P}Q = D = (d_{ij}\delta_{ij}) \quad (2.30)$$

Since \hat{P} is positive definite, the diagonal elements of D are positive.

Therefore the matrix T can be obtained from the formula

$$T = Q\left(\frac{\delta_{ij}}{\sqrt{d_{ij}}}\right) \quad (2.31)$$

The advantage of the above numerical technique is that by derivation of symmetric matrices it is possible to avoid the need for inversion of a general matrix which would tend to involve a large computational error.

Before proceeding with the numerical solution of Winters, the elements of the matrices P and \hat{P} must of course be estimated in terms of the probabilities that a document contains certain combination of

keywords as defined in Section 2.2.

2.4 Summary.

Use of the latent class concept, and the procedure for numerical solution of the equations as described above, appears attractive as a means of determination of document classes. The only probabilities required to be known are those that involve word associations within documents. It is not necessary to begin with a subdivision of documents into classes since this is determined as a result of the numerical solutions.

However, the above analysis is based on the assumption that disjoint sets of documents with the required latent class properties do, in fact, exist. It is also supposed that such classes, if they exist, have significance to users of the document data base.

If disjoint sets of documents with latent class properties do not exist for a given data base, the fact will be apparent in that the above procedure will not lead to a meaningful solution of the accounting equations. In order to be meaningful, a solution must lead to probability values that all lie within the range of 0 to 1. This condition is examined in the next Chapter.

CHAPTER III

APPLICATIONS OF LATENT CLASS ANALYSIS

3.1 Application of Winters' Method Using Experimental Data.

Winters did not perform any practical experiments to verify the applicability of his numerical solution to determine document classification. Instead, he gave an artificial example to illustrate the mathematical techniques when H , \hat{H} , and G are 4×4 matrices. The P and \hat{P} were computed from the relations $P = H'G\hat{H}H$ and $\hat{P} = H'GH$. Then, by application of his numerical techniques, he examined whether the original values resulted for H , \hat{H} , and G . In fact the computed values were in agreement with those assumed. This example only proved that his numerical techniques were valid, and that the equations did not become ill-conditioned for his example of 4×4 matrices. He made no attempt to find a solution of the equations that result for matrices of higher order or for matrices derived from real document data.

We have performed one experiment in which 7006 titles from the acoustic literature were used to compute the necessary probabilities for P and \hat{P} . In our experiment, the probability p_{ij} was computed as the probability that a document contains the same keyword K_i twice, and also the probability p_{ijN} was computed in the same manner. Details of the acoustics data base are given in Chapter V.

The following six words were selected arbitrarily as keywords:

1. ABSOR
2. EAR
3. NOISE

4. SPEEC
5. ULTRA
6. WATER

They are listed in truncated form to indicate the ABSOR might denote ABSORb, ABSORpion, etc., and similarly that NOISE might include NOISEs, etc.. The form of truncation is described further in Chapter V.

The matrices of probabilities P and \hat{P} were computed to give

$$P = \begin{bmatrix} 0.032258 & 0.002855 & 0.000143 & 0.001570 & 0.0 & 0.003283 \\ 0.002855 & 0.0 & 0.0 & 0.0 & 0.0 & 0.001142 \\ 0.000143 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.001570 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.003283 & 0.001142 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\hat{P} = \begin{bmatrix} 1.0 & 0.031402 & 0.006709 & 0.052098 & 0.021696 & 0.084499 \\ 0.031402 & 0.001285 & 0.0 & 0.0 & 0.0 & 0.011704 \\ 0.006709 & 0.0 & 0.0 & 0.000714 & 0.000285 & 0.0 \\ 0.052098 & 0.0 & 0.000714 & 0.001570 & 0.002997 & 0.000143 \\ 0.021696 & 0.0 & 0.000285 & 0.002997 & 0.000143 & 0.0 \\ 0.084499 & 0.011704 & 0.0 & 0.000143 & 0.0 & 0.000571 \end{bmatrix}$$

The next step was to derive the orthogonal matrix T which satisfies $T'\hat{P}T = I$. First, in order to determine the eigenvalues and eigenvectors of \hat{P} , Householder's method was applied to compute $Q'\hat{P}Q = D$ where the diagonal elements of the diagonal matrix D are the eigenvalues of \hat{P} . The

corresponding eigenvectors appear as the column vectors of Q . As a result of this computation it was found that three of the six eigenvalues were negative. This implies that the matrix \hat{P} is not positive definite. The computed eigenvalues and eigenvectors are as follows:

$$\lambda = \begin{matrix} 0.011309 & 0.008733 & 0.000015 & -0.000181 & -0.002923 & -0.013384 \end{matrix}$$

$$Q = \begin{bmatrix} -0.994416 & -0.041449 & -0.048236 & 0.035995 & 0.040402 & 0.064534 \\ -0.031884 & 0.686161 & 0.340707 & -0.252564 & -0.279383 & 0.519848 \\ -0.006639 & -0.072710 & 0.727169 & 0.679904 & 0.056715 & -0.020214 \\ -0.051388 & -0.401225 & 0.257334 & -0.256817 & -0.813892 & -0.204406 \\ -0.021490 & -0.247111 & 0.527305 & -0.634407 & 0.504591 & -0.057784 \\ -0.083511 & 0.547845 & 0.092443 & -0.064951 & -0.007965 & -0.824660 \end{bmatrix}$$

Since \hat{P} is not positive definite, we cannot proceed to the next stage of Winters' method to evaluate the matrix T .

The above example is not exceptional in producing a matrix \hat{P} that does not lead to determination of latent classes. Various choices of sets of keywords have been found to generally lead, either to a matrix \hat{P} that is not positive definite, or to determination of "probabilities" that do not all lie within the range 0 to 1.

However, even if a set of disjoint latent classes does not exist, there arises the question as to whether there exist classes that are almost disjoint, and for which the accounting equations may be approximately true. This is investigated in the next section.

3.2 An Attempt to Use Latent Class Analysis.

3.2.1 General.

Instead of attempting a matrix solution for latent class analysis, the present section presents a different method to determine latent classes and their associated probabilities.

In this new method determination of the number of latent classes is still a difficult problem. As in the method of Winters, we assume that the number of latent classes is equal to the number of keywords. There is justification for this assumption since in the special instance that no keywords tend to associate, then the number of latent classes is certainly equal to the number of keywords. Also, if the number of existing latent classes is, in fact, less than the number of keywords, then the probabilities corresponding to the non-existing latent classes will be computed as zeros, and the assumption will still be valid. The new numerical method will be called the "minimizing method".

3.2.2 Numerical Solution.

The original statement of the problem of latent class analysis involves solution of the set of equations defined in (2.4), (2.5), (2.6) (2.7) and so forth.

For practical application however, it is reasonable to make the following assumptions:

1. Significant associations of keywords within documents never involve sets of more than three keywords. This means that only p_i 's, p_{ij} 's, and p_{ijk} 's need be considered, but not $p_{ijk\ell}$'s, and etc..
2. If p_{ij} 's or p_{ijk} 's are sufficiently small, then the equations (2.6)

and (2.7) may be neglected.

In the case of N keywords and N latent classes, let us define a function $F(G,H)$ with $N(N+1)$ variables, $G = (g^\ell)$ and $H = (h_i^\ell)$, as follows:

$$\begin{aligned}
 F(G,H) = & \left(\sum_{\ell=1}^N g^\ell - 1 \right)^2 + \sum_{i=1}^N \left(\sum_{\ell=1}^N g^\ell h_i^\ell - p_i \right)^2 + \sum_{i=1}^N \sum_{j=1}^N \left(\sum_{\ell=1}^N g^\ell h_i^\ell h_j^\ell - p_{ij} \right)^2 \\
 & + \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \left(\sum_{\ell=1}^N g^\ell h_i^\ell h_j^\ell h_k^\ell - p_{ijk} \right)^2
 \end{aligned} \tag{3.1}$$

Obviously $F(G,H)$ is a non-negative function and if, and only if, each term of this function is equal to zero, then the function has a minimum value of zero. This minimum value occurs when the g 's and h 's correspond to the latent class structure that satisfies the equations (2.4), (2.5), (2.6), and (2.7).

It is obvious that the function $F(G,H)$ has concave form at the solution points, because the partial second derivatives of $F(G,H)$ with respect to the g 's and h 's are always positive.

For the purpose of the computations, new variables x_i^ℓ and y_{ij} and y_{ijk} are defined as follows:

$$x_i^\ell = h_i^\ell / p_i \tag{3.2}$$

$$y_{ij} = p_{ij} / p_i p_j \tag{3.3}$$

$$y_{ijk} = p_{ijk} / p_i p_j p_k \tag{3.4}$$

The function $F(G,H)$ may then be rewritten as the function

$$\begin{aligned}
F(G,X) = & \left(\sum_{\ell=1}^N g^{\ell} - 1 \right)^2 + \sum_{i=1}^N \left(\sum_{\ell=1}^N g^{\ell} x_i^{\ell} - 1 \right)^2 + \sum_{i=1}^N \sum_{j=1}^N \left(\sum_{\ell=1}^N g^{\ell} x_i^{\ell} x_j^{\ell} - y_{ij} \right)^2 \\
& + \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \left(\sum_{\ell=1}^N g^{\ell} x_i^{\ell} x_j^{\ell} x_k^{\ell} - y_{ijk} \right)^2
\end{aligned} \quad (3.5)$$

The first two sets of summations will be zero if g^{ℓ} and x_i^{ℓ} are chosen so that

$$g^N = 1 - \sum_{\ell=1}^{N-1} g^{\ell} \quad (3.6)$$

$$x_i^N = (1 - \sum_{\ell=1}^{N-1} g^{\ell} x_i^{\ell}) / g^N \quad (3.7)$$

where the subscript i varies from 1 to N .

If the conditions (3.6) and (3.7) are satisfied, then the function $F(G,X)$ to be minimized may be reduced to

$$\hat{F}(G,X) = \sum_{i=1}^N \sum_{j=1}^N \left(\sum_{\ell=1}^N g^{\ell} x_i^{\ell} x_j^{\ell} - y_{ij} \right)^2 + \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \left(\sum_{\ell=1}^N g^{\ell} x_i^{\ell} x_j^{\ell} x_k^{\ell} - y_{ijk} \right)^2 \quad (3.8)$$

Among the possible numerical techniques to solve the minimization problem, the method of steepest descent (21) is suggested. It may be described as follows. In the neighborhood of the solution, the function $\hat{F}(G,X)$, say $\hat{F}(Z)$, has a concave surface. If an initial value \underline{Z}_0 is chosen close to the solution of $\hat{F}(\underline{Z}) = 0$ then a better approximation \underline{Z}_1 is supposed in the form

$$\underline{Z}_1 = \underline{Z}_0 + \underline{\alpha}_0 d_0 \quad (3.9)$$

where $\underline{\alpha}_0$ forms a vector of step size, and d_0 indicates the direction of steepest descent. In general d_i is defined as $\text{grad } \hat{F}(\underline{Z}_i)$ so that (3.9)

becomes

$$\underline{Z}_{i+1} = \underline{Z}_i + \underline{\alpha}_i \text{ grad } \hat{F}(\underline{Z}_i) \quad (3.10)$$

In the computation, all elements of the vector $\underline{\alpha}_i$ are assumed to be constant, and coordinate axes are used in place of $\text{grad } \hat{F}(\underline{Z}_i)$.

For the first approximation to the solution of $\hat{F}(G,X) = 0$, the following values can be chosen.

$$g^\ell = \frac{1}{N} \quad (\ell=1,N) \quad (3.11)$$

$$x_i^\ell = 1 \quad (\ell=1,N \text{ and } i=1,N) \quad (3.12)$$

which satisfy the relations in (3.6) and (3.7).

In the iteration procedure, the next approximates are calculated by adding a small perturbation $\underline{\alpha}$ or $-\underline{\alpha}$ to the previous approximations so that $\hat{F}(G,X)$ can be decreased. The iteration will be repeated until the value of $\hat{F}(G,X)$ becomes sufficiently small or until a given number of iterations is completed.

It should be noted that the method described may lead only to approximate solutions of the latent class equations since, in fact, it may happen that no real solutions exist. However, approximate solutions may be quite satisfactory in practice since it is not a serious problem if there is slight overlap between the different classes of documents.

3.2.3 Application of Proposed Method.

Using the numerical techniques stated in the previous section, a sample computation was performed. Given a sample solution for 6 keywords and 6 classes, the probabilities were computed backwards. The sample solution and modified probabilities were chosen in Table 3.1.

Table 3.1 A Sample Solution of the Accounting Equations

$g^1 = 0.05$	$x_1^1 = 0.307692$	$x_2^1 = 1.142858$	$x_3^1 = 0.923077$	$x_4^1 = 0.895523$	$x_5^1 = 1.403508$	$x_6^1 = 1.212122$
$g^2 = 0.10$	$x_1^2 = 0.923077$	$x_2^2 = 2.285715$	$x_3^2 = 0.307692$	$x_4^2 = 0.597015$	$x_5^2 = 1.754387$	$x_6^2 = 1.212122$
$g^3 = 0.15$	$x_1^3 = 0.615385$	$x_2^3 = 0.571429$	$x_3^3 = 0.615385$	$x_4^3 = 1.194030$	$x_5^3 = 1.052631$	$x_6^3 = 1.818183$
$g^4 = 0.20$	$x_1^4 = 0.615385$	$x_2^4 = 1.142858$	$x_3^4 = 1.230768$	$x_4^4 = 1.492537$	$x_5^4 = 0.350877$	$x_6^4 = 1.212122$
$g^5 = 0.20$	$x_1^5 = 1.538462$	$x_2^5 = 1.142858$	$x_3^5 = 0.615385$	$x_4^5 = 0.298508$	$x_5^5 = 1.052831$	$x_6^5 = 0.606061$
$g^6 = 0.30$	$x_1^6 = 1.230768$	$x_2^6 = 0.571429$	$x_3^6 = 1.538462$	$x_4^6 = 1.194030$	$x_5^6 = 1.052631$	$x_6^6 = 0.606061$

The resulting modified probabilities y_{ij} 's and y_{ijk} 's are

$y_{12} = 0.984160$	$y_{13} = 1.008284$	$y_{14} = 0.895522$	$y_{15} = 1.036437$	$y_{16} = 0.857809$
	$y_{23} = 0.861539$	$y_{24} = 0.904051$	$y_{25} = 1.072681$	$y_{26} = 1.021645$
		$y_{34} = 1.125143$	$y_{35} = 0.917679$	$y_{36} = 0.913753$
			$y_{45} = 0.900759$	$y_{46} = 1.067390$
				$y_{56} = 0.988866$
$y_{123} = 0.827726$	$y_{124} = 0.771527$	$y_{125} = 1.091961$	$y_{126} = 0.884449$	
	$y_{134} = 0.058376$	$y_{135} = 0.979965$	$y_{136} = 0.797562$	
		$y_{145} = 0.857252$	$y_{146} = 0.829420$	
			$y_{156} = 0.883328$	
	$y_{234} = 0.928982$	$y_{235} = 0.777328$	$y_{236} = 0.831169$	
		$y_{245} = 0.825946$	$y_{246} = 0.992441$	
			$y_{256} = 1.099719$	
		$y_{345} = 0.953935$	$y_{346} = 1.743448$	
			$y_{356} = 0.798266$	
			$y_{456} = 0.939482$	

where the modified probabilities y_{ij} 's and y_{ijk} 's are defined only for $i \neq j$, $i \neq k$, and $j \neq k$.

Substituting the initial approximation $g^{\ell} = 1/6$ and $x_i^{\ell} = 1$, the function $\hat{F}(G,X)$ was found to have a value of 6.259737. After 2550 iterations, the value of $\hat{F}(G,X)$ was reduced to 0.005164, and the resulting values of g 's and x 's are as shown in Table 3.2.

The labelling of the latent classes of this approximation is, of course, not necessarily in the same order as those of Table 3.1. Comparison of the two tables indicates that g^1 of Table 3.2 corresponds to g^2 of Table 3.1, g^2 to g^6 , g^3 to g^1 , g^4 to g^5 , g^5 to g^3 , and g^6 to g^4 . The correspondence is shown in Table 3.3.

It is apparent from Table 3.3 that the solution is considerably different from the exact values. In order to increase the accuracy, more iterations are needed. However, this is not an easy task, because at a high number of iterations the value of $\hat{F}(G,X)$ is apt to oscillate and does not converge smoothly. The rate of convergence, and tendency to oscillate, is dependent on the choice of the constant scaling factor which denotes the step size for the next iteration. Therefore, as the iterations proceed, in order to improve the rate of convergence the value of the step size must be suitably changed as necessary.

3.2.4 Discussion.

The sample calculations of the previous section illustrates that, even if estimations of step size are made at certain stages, the iterations must be repeated many times in order to obtain a solution with acceptable accuracy. This problem may not be very serious in the sample calculations which applied to only 6 keywords and 6 latent classes. However, in any practical instance in which there may be hundreds of keywords, the method involves a lot of multiplications to

Table 3.2 Approximates Given by Minimizing Method

$g^1 = 0.07$	$x_1^1 = 0.107859$	$x_2^1 = 1.097119$	$x_3^1 = 0.337107$	$x_4^1 = 1.183116$	$x_5^1 = 1.143532$	$x_6^1 = 2.323439$
$g^2 = 0.28$	$x_1^2 = 1.201924$	$x_2^2 = 1.408411$	$x_3^2 = 0.448426$	$x_4^2 = 0.347828$	$x_5^2 = 1.327293$	$x_6^2 = 0.790367$
$g^3 = 0.06$	$x_1^3 = 0.823315$	$x_2^3 = 0.885635$	$x_3^3 = 0.040945$	$x_4^3 = 1.150599$	$x_5^3 = 0.740308$	$x_6^3 = 1.417387$
$g^4 = 0.23$	$x_1^4 = 0.994804$	$x_2^4 = 1.172207$	$x_3^4 = 1.151906$	$x_4^4 = 1.193593$	$x_5^4 = 0.978572$	$x_6^4 = 1.123346$
$g^5 = 0.18$	$x_1^5 = 0.643038$	$x_2^5 = 0.710825$	$x_3^5 = 1.699266$	$x_4^5 = 1.525926$	$x_5^5 = 0.487116$	$x_6^5 = 0.969346$
$g^6 = 0.19$	$x_1^6 = 1.375196$	$x_2^6 = 0.420043$	$x_3^6 = 1.439788$	$x_4^6 = 1.025659$	$x_5^6 = 1.110025$	$x_6^6 = 0.534207$

Table 3.3 Comparison of Postulated g^λ and x_i^λ with Computed Values (in parentheses)

	g^λ	x_i^λ					
		i=1	i=2	i=3	i=4	i=5	i=6
$\lambda = 1$	0.05 (0.06)	0.31 (0.82)	1.14 (0.89)	0.92 (0.04)	0.90 (1.15)	1.40 (0.74)	1.21 (1.42)
$\lambda = 2$	0.10 (0.07)	0.92 (0.11)	2.29 (1.10)	0.31 (0.34)	0.60 (1.18)	1.75 (1.14)	1.21 (2.32)
$\lambda = 3$	0.15 (0.18)	0.62 (0.64)	0.57 (0.71)	0.62 (1.70)	1.19 (1.53)	1.05 (0.49)	1.82 (0.97)
$\lambda = 4$	0.20 (0.19)	0.62 (1.38)	1.14 (0.42)	1.23 (1.44)	1.49 (1.03)	0.35 (1.11)	1.21 (0.53)
$\lambda = 5$	0.20 (0.23)	1.54 (0.99)	1.14 (1.17)	0.62 (1.15)	0.30 (1.19)	1.05 (0.98)	0.61 (1.12)
$\lambda = 6$	0.30 (0.28)	1.23 (1.20)	0.57 (1.41)	1.54 (0.45)	1.19 (0.35)	1.05 (1.33)	0.61 (0.79)

compute the function $\hat{F}(G,X)$ so that the accuracy of the approximated values is doubtful. Furthermore, the enormous number of iterations that must be performed make the method very expensive in computer time.

3.3 Conclusions Regarding the Limitations, or Unsuitability, of Latent Class Determination.

3.3.1 Winters' Method.

Two different methods have been described and applied to the solution of the system of non-linear equations which define the latent class structure. First, Winters' numerical technique was applied directly to determine the latent classes for an existing data base of acoustic literature. The method was not successful because the assumption that \hat{P} is positive definite did not hold in this case. In fact \hat{P} had three positive eigenvalues and three negative eigenvalues. In order to obtain T it is necessary to compute the square roots of the eigenvalues of \hat{P} , and if these eigenvalues are negative, it is impossible to obtain a real T . The definiteness of P was not checked, but P might not be positive definite either. If, and only if, the conditions (2.15) and (2.16) hold, then the Winters' techniques can be applied. However an arbitrary data base does not in general satisfy these conditions, and so a latent class structure does not generally exist. This implies that, in general, keywords do not occur statistically independently in each class, and so $h_{ij}^{\ell} \neq h_i^{\ell} h_j^{\ell}$.

An attempt to use the Winters' method while avoiding the above difficulties might proceed as follows:

1. Select a few keywords arbitrarily.

2. Compute P and \hat{P} .
3. If both P and \hat{P} are positive definite, then add more keywords and go to 2. Otherwise, reject some keywords and add others, and go to 2.

The above step might be continued until realisable latent class structure results. Of course, even if such a procedure is applicable to a practical data base that involves hundreds of keywords, it may be very time consuming and expensive. Our investigations have provided no evidence to suggest the practical feasibility of determining the latent class structure of a data base that contains several hundreds of keywords. There is also an inconsistency between the definitions and the numerical approach by Winters. Recalling the definitions of latent class structure in (2.5), (2.6), and (2.7), the subscripts i , j , and k are defined to have unequal values and so the p_{ij} 's and p_{iN} 's are undefined. However these undefined terms do appear as diagonal elements of the matrices P and \hat{P} , and hence direct application of the Winters' analysis is not possible. Winters did not mention this fact in his paper (24). This is, however, a minor problem, and may be overcome by changing the definition such that the subscripts i , j , and k may be the same. The probability p_{ij} may be defined as the probability that a document contains at least two i^{th} keywords, and the probability p_{iN} as the one that a document contains at least two i^{th} and one N^{th} keywords. This change of definition does not destroy the latent class structure, because it is not irrational to apply the independence assumption to h_{ij}^{ℓ} 's and h_{iN}^{ℓ} 's such that $h_{ij}^{\ell} = h_i^{\ell} h_j^{\ell}$ and $h_{iN}^{\ell} = h_i^{\ell} h_N^{\ell}$.

According to the formula (2.9) which evaluates the ordering ratio, every document is classifiable, even ones with no keywords. This appears

to be at variance with the intuitive idea that the absence of keywords implies no information, and hence such documents cannot be classified.

We suggest one alternative to the above. Change the formula (2.9) which calculates the ordering ratio, to the form

$$p^{\ell} = \frac{g^{\ell} h_1^{\ell} h_2^{\ell} \dots h_M^{\ell}}{\sum_{k=1}^L g^k h_1^k h_2^k \dots h_M^k} \quad (3.13)$$

which neglects the non-existing keywords. This formula is more easily computed than the one in (2.9), since formula (3.13) requires only $M(N+1)$ multiplications where M is the number of keywords in a given document. In contrast, the formula (2.9) requires $N(N+1)$ multiplications for every case.

3.3.2 Minimizing Method.

The other approach proposed in Section 2 to solve latent class equations is the minimizing method in which the necessary probabilities are computed to minimize the positive function $\hat{F}(G, X)$. The sample computations were for 6 keywords and 6 latent classes. Even after 2550 iterations, with about 14 minutes execution time, the approximation was not close to the exact solution. Thus, the iterative procedure does not provide an economic solution to the problem.

3.3.3 Summary.

A fundamental question concerning latent class analysis is whether there exist such latent classes for an arbitrary group of documents. Since the required probabilities are estimated with possibility of some numerical error because of finite sampling, it is very unsatisfactory

to have latent class determination dependent on methods whose results are affected by small changes in the numerical data.

In general, the latent class analysis involves too many unknowns, g 's and h 's, and it requires a large system of non-linear equations. In fact, solution of such equations poses a problem in numerical analysis, and it is not clear that the resulting latent classes are sufficiently well-defined to be useful in document classification.

CHAPTER IV

ATTRIBUTE ANALYSIS

4.1 Classification by Attribute Number.

On the basis of Luhn's pioneer work (10, 11), in 1961 M. E. Maron applied statistical techniques to the problem of automatic classification. He derived a formula based on probabilities of word occurrences and subject categories. He used the computer to evaluate the probability that a document which contains a certain combination of keywords also belongs to a certain category. In addition to developing prediction formulas based on probabilities, he carried out experimental work which may be used as the basis to determine the direction of further studies.

Maron's prediction formula for classification is based entirely on the statistical associations between categories and certain keywords in documents. Suppose that a document contains only one keyword K_i . Then the probability that this document belongs to the k^{th} category C_k is expressed by

$$P(C_k; K_i) = \frac{P(K_i; C_k)P(C_k)}{P(K_i)} \quad (4.1)$$

where $P(K_i; C_k)$ is the probability that a document in the k^{th} category C_k contains the i^{th} keyword K_i . The term $P(C_k)$ is the probability that a document is in the category C_k , and the term $P(K_i)$ is the probability that a document contains the keyword K_i .

The value of $P(C_k; K_i)$ indicates the degree of association of the given document with the k^{th} category. Therefore, if regarded as a function of C_k , the function $P(C_k; K_i)$ has its largest value at $k = 9$, then

the 9th category is the most suitable category for the document.

More generally, suppose that a document has M keywords $\{K_i\}_1^M$. The probability that the document belongs to the category C_k is then

$$P(C_k; \{K_i\}_1^M) = \frac{P(\{K_i\}_1^M; C_k)P(C_k)}{P(\{K_i\}_1^M)} \quad (4.2)$$

Since $P(\{K_i\}_1^M)$ is independent of the choice of the categories the above expression may also be written in the form

$$P(C_k; \{K_i\}_1^M) = kP(\{K_i\}_1^M; C_k)P(C_k) \quad (4.3)$$

where k is a constant which is independent of the choice of categories.

In order to simplify the computations Maron made the important assumption that in each category the keywords occur in a statistically independent manner. Then (4.3) may be further simplified to become

$$P(C_k; \{K_i\}_1^M) = kP(C_k) \prod_{i=1}^M P(K_i; C_k) \quad (4.4)$$

and $P(C_k; \{K_i\}_1^M)$ is then called an "attribute number". π signifies the multiplication of terms as i ranges from 1 to M.

4.2 Selection of Data.

In his research Maron chose an experimental data base of some 405 abstracts chosen from the computer journal literature. He attempted to classify the documents by automatic processing of the abstracts. These abstracts are in the IRE Transactions on Electronic Computers, vol. EC-8, no. 1, published by the IRE Professional Group on Electronic Computers in 1959.

The 405 abstracts were divided into two groups. Group 1 consisted of 260 abstracts which were published in the March and June issues of

1959. Group 2 consisted of 145 abstracts which were available in the September issues of 1959. Group 1 formed the data base for the computation of the statistical values required for the automatic classification. Group 2 was used to test the theory of the category prediction based on use of the statistical data collected from Group 1. Therefore Group 2 was not considered until all the statistical procedures had been performed on Group 1. The 260 abstracts in Group 1 contained more than 20,000 words, 3,263 different words, and the average number of words in a document abstract was 79.

4.3 Selection of Categories.

The IRE had its own categories for the classification of computer literature. They consisted of 10 categories and about 15 subcategories. However, Maron considered that the IRE categories were not distinct enough to be used as a test of his procedures, and so he grouped the documents among 32 subject categories. The 260 documents of Group 1 were then manually classified into the supposed proper categories. Most of the documents fell naturally into a single category, but about 20% of the documents belonged to two categories, and some of them belonged to three categories.

4.4 Selection of Keywords.

In his work Maron used 90 keywords and formulated a theoretical analysis to relate documents and keywords as described below.

According to Shannon's theory of entropy, the average uncertainty H with which a document may be assigned to a category is

$$H = - \sum_{k=1}^{32} P(C_k) \log_2 P(C_k) \quad (4.5)$$

where $P(C_k)$ denotes the probability that a document belongs to the k^{th} category.

Suppose that a document is indexed by one word, say W_i . Then the average uncertainty H_i that the document belongs to any one of the 32 categories can be represented by

$$H_i = - \sum_{k=1}^{32} P(C_k; W_i) \log_2 P(C_k; W_i) \quad (4.6)$$

where $P(C_k; W_i)$ is the probability that a document keyworded by the word W_i belongs to the k^{th} category C_k .

Since the difference $H - H_i$ is the uncertainty removed by the selection of the word W_i as a keyword, the keywords should be decided by computing $H - H_i$ for all words that appear on the data, and by ranking the resulting values in decreasing order. Such a list then shows the order of efficient keywords.

However, Maron did not follow this method to determine the 90 keywords from Group 1. Instead, he first removed the 55 function words (e.g. the, of, a, etc.) which had a total of 8,402 occurrences. Thus, about 2% of the different words accounted for over 40% of the total occurrences. He also removed frequently occurring words (e.g. computer, data, system, etc.), and the 2,120 rarely occurring words used only once or twice and which accounted for 65% of the total 3,263 different words. About 1,000 different words remained as possible keywords. Among them, 90 words were each found to occur predominantly in a single category, and were considered to be suitable choices for keywords for automatic classification.

In the present thesis, which describes techniques applied to a data base formed from the acoustic literature, we do not follow the

method of keyword selection used by Maron. Our method is described in detail in Chapter V.

4.5 Experimental Results.

Maron's experiment was divided into two separate parts. The first was applied to documents of Group 1. The second was applied to documents of Group 2. As previously stated, all necessary data (90 keywords and the value of $P(C_k)$'s and $P(K_i;C_k)$'s) was determined by use of documents of Group 1. Therefore, the results of Group 2 provided information for discussion of the generality of Maron's method and suggestion of future extension of work in automatic classification.

As may be seen from the prediction formula (4.4), if at least one of the numbers $P(K_i;C_k)$ is zero, then the attribute number becomes zero. In order to avoid this disadvantage, Maron assigned a very small value (viz. 0.001) to replace the zero values of the $P(K_i;C_k)$. This technique proved very useful in the classification of Group 2. Because the values of $P(K_i;C_k)$ were computed only from documents of Group 1, some of the $P(K_i;C_k)$ needed to compute the attribute numbers in Group 2 were not available but were approximated by the assumed small value.

The results obtained by Maron are summarized in Table 4.1. It may be noted that for the documents of Group 1, the attribute analysis method worked just as well as the manual judgments. Of the 247 documents available for automatic classification there were 209 for which the computer correctly assigned the largest attribute number. As described in Section 4.3, the manual examinations could not place 20% of the documents into just one category. This suggests that about 20% of uncertainty is likely to be involved in any type of classification of those

documents. Thus, it may be regarded as surprisingly good that 84.6% of documents in Group 1 were classified under correct categories by the computer. In his paper Maron did not give complete details regarding classification of Group 2. Thus we cannot discuss his results precisely. However, the figures available in Table 4.1 for documents that contain more than one keyword show that 44 out of 85 documents were correctly classified. Thus approximately 50% of documents were correctly classified under only one category.

4.6 Summary.

The value of Maron's pioneer work on automatic classification is, not only that he used a statistically based classification system to successfully derive the proper category for a document, but that he also introduced the concept of "attribute number" to describe the degree of association between a given document and category. Maron made the statement that "...., instead of stating that either a document belongs to a given category or not, it would be more realistic to recognize that a document can belong to a category to a degree (i.e., with a weight)." The degrees are, in fact, indicated by the set of attribute numbers.

In summary, the experimental results of Maron are sufficiently encouraging for us to proceed to modify, and attempt to improve, Maron's method of attribute analysis which forms the foundation of a statistical approach to relationships between keywords and categories.

It should be noted that the keywords used by Maron were chosen from the words that appeared in the documents abstracts. An automatic choice of keywords therefore requires that the abstracts be available in machine

Table 4.1 Summary of Maron's Results

	Group 1	Group 2
Total number of documents	260	145
Number of documents with no keyword	12	20
Errors during processing	1	0
Number of documents available for automatic classification	247	125
Number of documents with only one keyword	37	40
Correct classifications	18	-
% of N_1	48.7%	-
Number of documents with more than one keyword	210	85
Correct classifications	191	44
% of N_2	91.0%	51.8%
Total number of correct classifications	209	-
% of N_k	84.6%	-
% of N_t	80.4%	-

readable form. Computer processing of abstracts is, of course, more costly than similar processing of document titles, and there arises the question as to whether an efficient choice of keywords could be based on processing of title words only. This is one of the questions considered in the subsequent chapters.

CHAPTER V

ACOUSTICS DATA BASE AND SELECTION OF KEYWORDS

5.1 Selection of Data.

The data base used to perform our experiments consists first of 1572 titles of papers published in the Journal of the Acoustical Society of America (JASA) in 1966, 1967, 1968, and 1961.

An individual datum consisting of journal name, year, volume number, page, authors, and title is punched on cards to provide the data base accessible to the computer. The author and title words are truncated to five letters. A detailed description of this acoustics data is given in JASA vol. 43, no. 6 (7). Although truncation might be undesirable in a data base used for information retrieval, it has no effect on the validity of the results of the present thesis.

The 1572 titles are divided into four groups. Group 1 consists of 395 titles which were all published in 1966. Group 2 consists of 385 titles which were published in 1967. Group 3 consists of 506 titles which were published in 1968. Group 4 consists of 286 titles which were published in 1961. Group 1 will be used as the base data for choice of 200 keywords and to estimate probabilities, and necessary values, required for the experimental model. The automatic classification schemes will be tested over group 1, group 2, group 3, and group 4 separately, and the results will be compared with those of previous researchers.

5.2 Selection of Categories.

JASA has prepared 16 main subject categories to classify articles

issued by JASA. Each of the 16 categories has been further divided into several sub-categories.

In our experimental investigation the JASA sub-categories are not used because they are too precise to distinguish concepts of articles. Furthermore, of the 16 main categories, the categories 1, 3, and 8 are not used because very few articles have been issued in these subject categories. Thus, 13 main categories out of 16 are used in our experiments. An additional category is provided to classify articles which cannot be classified under either of the 13 categories. Thus 14 main categories are renumbered and are as listed below:

1. Architectural Acoustics.
2. Physiological and Psychological Acoustics.
3. Acoustical Instruments and Apparatus.
4. Music and Musical Instruments.
5. Noise and Noise Control.
6. Speech Communication.
7. Ultrasonics.
8. Radiation and Scattering.
9. Mechanical Vibrations and Shock.
10. Underwater Sound.
11. Aeroacoustics, Macrosonics.
12. Acoustic Signal Processing.
13. Bioacoustics.
14. Miscellaneous.

The titles used in the experiment are manually classified following the above classification schedule accepting the JASA assignment of category, and the indication of subject category is punched on each data

card that describes the document.

5.3 Selection of Keywords.

M.E. Maron suggested a method to select keywords for the classification system. The method is based on Shannon's theory of entropy. The detailed description of this method is stated in Chapter IV. The direct application of Maron's suggestion to the keyword selection, however, raises a problem which may be illustrated as follows.

According to Maron's theory, the uncertainty of the correct classification of each word of a document has its minimum value of 0 if, and only if, the word occurs only in documents of a particular subject category. This theory completely neglects the frequency of the word occurrence. For example, two different words, W_i and W_j , which occur 10 times and 20 times respectively in documents, may have the same zero value of the uncertainty, because they each occur in documents of a particular category. However, the word W_i will classify 10 documents correctly, while the word W_j will similarly classify 20 documents. Therefore the word W_j should be considered to be the better indication of subject category than the word W_i .

In the present section the method of keyword selection emphasizes the degree of accuracy of the total automatic classification system. For documents that contain the word W_i , let $N(C_k, W_i)$ denote the number of documents that should be classified in category C_k . Then $\sum_{t \neq k} N(C_t, W_i)$ is the number of documents classified under a category C_t different from C_k .

First suppose that only one keyword is used to index each document. To place every document indexed by word W_i under category C_k produces

$N(C_k, W_i)$ correct document classifications but produces $\sum_{t \neq k} N(C_t, W_i)$ incorrect classifications. The value $N(C_k, W_i) - \sum_{t \neq k} N(C_t, W_i)$ is the difference between the number of correct classifications and the number of incorrect classifications, and it is therefore a measure of the appropriateness of the word W_i as a single keyword to describe the class C_k .

In the selection of keywords, the difference $N(C_k, W_i) - \sum_{t \neq k} N(C_t, W_i)$ should be made as large as possible. Of course even to make this value positive is not always possible. For example, some very common words such as ACOUSTIC and SOUND, etc. in the present data are distributed approximately uniformly throughout all 14 categories, and so their differences may be negative. Thus, commonly occurring words will tend to be automatically eliminated from consideration as keywords.

Each selected keyword W_i should make the function $F(C_k) = N(C_k, W_i) - \sum_{t \neq k} N(C_t, W_i)$ have a sharply defined peak at some C_k .

The proposed method of selection of keywords may be illustrated by reference to Table 5.1 which indicates the frequencies of words in categories for 6 keywords and 14 categories.

The first column of Table 5.1 indicates that documents containing the keyword ACOUS occur more frequently in C_{10} than in any other category. Thus, if all documents that contain ACOUS are to be assigned to a single category, then the category should be chosen as C_{10} . Of the 42 documents that contain ACOUS, this classification will classify 12 documents correctly and 30 incorrectly. Therefore the number of correct classifications exceeds the number of incorrect classifications by -18. Similarly, the second column of Table 5.1 shows that documents containing the keyword BINAU are most frequently in category C_2 .

Table 5.1 Word Frequency Table Used for Keyword Selection

	ACOUS	BINAU	DEEP	SOUND	VIBRA	WIDE
C_1	2	0	0	2	0	0
C_2	5	9	0	7	1	0
C_3	1	0	0	2	1	0
C_4	0	0	0	2	1	0
C_5	0	0	0	0	0	0
C_6	2	0	0	4	0	0
C_7	7	0	0	6	2	1
C_8	8	0	0	8	2	0
C_9	0	0	0	0	28	0
C_{10}	12	0	6	17	0	0
C_{11}	2	0	0	7	1	0
C_{12}	2	0	0	0	0	0
C_{13}	0	0	0	0	1	0
C_{14}	1	0	0	2	0	0
Correctly	12	9	6	17	28	1
Incorrectly	30	0	0	40	9	0
Difference	-18	9	6	-23	19	1

The final row of Table 5.1 suggests that keywords BINAU, DEEP, VIBRA, are better category indicators than is WIDE. Also, the keywords ACOUS and SOUND are poor choices of keywords for indication of category.

The above procedure was used to choose keywords from the 395 acoustic titles from 1966. The titles and categories were input through a computer program which computed the difference values as in Table 5.1 and then selected the keywords that corresponded to the 200 largest differences. The resulting keywords are listed in Appendix A. Author names, as well as title words, were allowed as keywords since it was not wished to exclude the possibility that certain author names might be very indicative of subject matter.

5.4 Statistics on Data.

The attributes of group 1 form the basis for the prediction about the attributes of group 2 and group 3. The statistical nature of group 1 is described below.

The 395 titles in group 1 contain a total of 3,231 words, and the average number of words per title is about 8.2. There are 1,327 different words contained in the titles and therefore each word occurs, on the average, in two or three titles.

The titles in group 1 were pre-classified under 14 subject categories as summarised in Table 5.2.

In Table 5.2, the large numbers that appear under categories 2, 7, 8, 9, and 10 indicate that in 1966 there were many papers that related to these particular five subject fields. It is interesting to list the similar statistics for group 2 in order to see the changes in research interest. The figures of group 2 are shown in Table 5.3.

Comparing the figures in Table 5.2 and Table 5.3, it is noticed

Table 5.2 Distribution of Number of Titles in Group 1 (1966) over 14 Categories.

Categories	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of titles	7	89	15	7	15	26	40	40	58	62	16	9	4	7
%	1.8	22.5	3.8	1.8	3.8	6.6	10.1	10.1	14.7	15.7	4.1	2.3	1.0	1.8

Table 5.3 Distribution of Number of Titles in Group 2 (1967) over 14 Categories.

Categories	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of titles	5	75	23	11	6	28	61	39	38	60	7	12	4	16
%	1.3	19.5	6.0	2.9	1.6	7.3	15.8	10.1	9.9	15.6	1.8	3.1	1.0	4.2

that in 1967 subject fields 3 and 7 which are Acoustical Instruments and Apparatus and Ultrasonics were becoming more popular, but subject fields 5, 9, and 11 which are Noise and Noise Control, Mechanical Vibration, and Shock and Aeroacoustics Macrosonics were becoming relatively less popular than in 1966. This fact provides a warning that in attribute analysis it may be very dangerous to use a partial set of articles as a base data to predict the attributes of the whole data.

There were 82 titles in group 1, 102 titles in group 2, 143 titles in group 3, and 96 titles in group 4 that did not have any of the 200 selected keywords, and the rest of the titles contained at least one, and up to six, keywords. The Table 5.4 gives the figures regarding the number of keywords in titles.

From Table 5.4, it follows that each title in group 1 contains an average of 1.8 keywords, each title in group 2 and group 3 contains an average of 1.3 keywords, and each title in group 4 contains an average of 1.2 keywords.

Table 5.4 Number of Keywords in Titles.

Number of Keywords	Group 1	Group 2	Group 3	Group 4
0	82	102	143	96
1	100	147	184	99
2	103	91	117	53
3	61	25	42	29
4	26	16	13	8
5	11	2	5	1
6	12	2	2	0
Total of Keyword Occurrences	720	490	633	329

CHAPTER VI

APPLICATION OF MARON'S ATTRIBUTE ANALYSIS TO ACOUSTICS DATA BASE

6.1 Experimental Results on Acoustic Data.

In Chapter V we have described in detail the acoustics data base and the 14 categories available for experimentation. A method of selection of keywords was described. This is the method used to choose the 200 keywords referred to in the present, and subsequent, chapter.

Following Maron's scheme, the 395 titles in the 1966 issues, the 385 titles in the 1967 issues, the 506 titles in the 1968 issues, and the 286 titles in the 1961 issues, are disjoined to form group 1, group 2, group 3, and group 4, respectively.

As shown in Table 6.1, in group 1 there were 82 out of 395 titles which did not contain any of the chosen 200 keywords; therefore automatic classification could not be undertaken for these 82 titles. At least one keyword appeared in the remaining 313 titles which were therefore regarded as suitable for classification by the method of Maron. For the titles that contained only one keyword the automatic classification process predicted the correct categories in 79 instances. The remaining 213 of the 313 titles contained more than one keyword, and exactly 191 titles were classified correctly. Thus, 270 out of 313 titles were automatically given correct categories and so for group 1 the accuracy was about 86.3%.

In group 1 the titles with at least one keyword were classified correctly with the quite high degree of accuracy of 79% and 89.7% respectively. This fact is not surprising when it is recalled that all necessary statistical data was computed on the basis of group 1. In

Table 6.1 Experimental Results of Maron's Method Applied to Acoustics Data.

	group 1 (1966)	group 2 (1967)	group 3 (1968)	group 4 (1961)
Total number of titles	395	385	506	286
Number of titles with no keyword	82	102	143	96
Number of titles with at least one keyword	313	283	363	190
Number of titles with only one keyword	100	147	184	99
Number of correct classifications	79	80	99	40
% of N_1	79.0%	54.4%	53.8%	40.4%
Number of titles with more than one keyword	213	136	179	91
Number of correct classifications	191	96	117	59
% of N_2	89.7%	70.6%	65.4%	64.8%
Total number of correct classifications	270	176	216	99
% of N_k	86.3%	62.2%	59.5%	52.1%
% of N_t	68.4%	45.7%	42.7%	34.6%

groups of 2 to 4, however, automatic classification gave a poor prediction (54.4%, 53.8%, and 40.4%, respectively) of correct categories for the titles with only one keyword. On the other hand, for the titles with more than one keyword the classification was relatively good, with an accuracy of 70.6% in group 2, 65.4% in group 3, and 64.8% in group 4. Therefore, for titles in which more than one keyword is used to index, it appears that a high degree of automatic classification may be achieved.

6.2 Discussion of Results.

We have described two experiments which have been performed in order to analyze Maron's automatic classification procedure. The first, described in Chapter IV, used the abstracts of the IRE Transactions on Electronic Computers; the other used titles from the Journal of the Acoustical Society of America. We cannot expect similar results from both experiments because of the differences in the type of data (one comprised abstracts, the other comprised titles), the methods of keyword selection and the category selection. However, the availability of two separate sets of results does provide more ground for evaluation of Maron's theory than would one alone.

Comparison of the group 1 results in Tables 4.1 and 6.1 shows that the present method of choosing keywords as described in Chapter V is very effective in that for the documents that contain only one keyword there are 79.0% that are classified correctly. In contrast, Maron's choice of keywords led to a correct classification of only 48.7% of such documents. It is interesting to note that the numbers of correct classifications of indexed documents (% of N_k) are 84.6% and 86.3%; hence one may conclude that for the acoustics data base the

document titles provide a satisfactory source of keywords.

Comparison of results for groups 1 to 4 in Table 6.1 shows that, while the number of correct classifications is less for group 2 to 4 than for group 1, the number does not change significantly with increase in the time interval between groups. This suggests that the vocabulary of significant title words does not change appreciably from year to year over an eight year period.

The 20 to 25% reduction in correct classification of documents not contained in group 1, whether they contain one or more keywords, suggests that the classification errors are caused by false initial classification or unsuitable titling of the base documents.

6.3 Possible Improvements in Procedure.

Maron suggested four methods by which his prediction procedure might be improved. The first way is to use more documents in group 1 in order to collect more stable statistical data. The second way is to increase the total number of keywords available for the classification. The third way is to apply more accurate calculation of the statistical terms; for example in order to predict $P(C_k; K_i, K_j)$ one might use $P(C_k)P(K_i; C_k)P(K_j; K_i, C_k)$ instead of $P(C_k)P(K_i; C_k)P(K_j; C_k)$ which is based on an assumption of independence of certain probabilities. The fourth way is to give more consideration to the frequency of occurrence of keywords in documents.

The first, the second, and the fourth methods appear likely to be profitable, because more data and keywords lead to more accurate classification statistics. However, there are two reasons why we cannot agree completely with Maron's third suggestion. One is that implementation requires a large computer memory to store more accurate

statistics. The other is that, although logically the direct computation of $P(C_k; K_i, K_j) = P(C_k)P(K_i; C_k)P(K_j; K_i, C_k)$ instead of the approximate $P(C_k)P(K_i; C_k)P(K_j; C_k)$ should lead to a better classification in group 1, it is doubtful whether the same is true for the other groups because there exists some bias between the groups. For the groups considered the experimental results suggest that there is no serious error caused by the assumption that in any category the keywords occur statistically independently.

In summary, attribute analysis for automatic classification seems to work fairly well. We believe, moreover, that the method is very satisfactory for documents with more than one keyword. It is less satisfactory for documents with only one keyword. Therefore, there is a need to derive a method suitable, not only for documents with several keywords, but also for ones with only one keyword. This is discussed in the next Chapter.

CHAPTER VII

MODIFIED ATTRIBUTE ANALYSIS

7.1 Maximization of Correct Document Classifications.7.1.1 Classification System.

The present section describes a classification system which attempts to maximize correct document classifications. The basic theory is similar to that for the keyword selection as described in Chapter V.

Suppose that a document is indexed by a set of M keywords, denoted by $\{K_i\}_1^M$. Of all documents indexed by $\{K_i\}_1^M$ let $N(C_k, \{K_i\}_1^M)$ be the number in category C_k . Obviously, for a document indexed by $\{K_i\}_1^M$, the category in which $N(C_k, \{K_i\}_1^M)$ has the largest value is the best one in which to classify the document.

However, if all possible values of $N(C_k, \{K_i\}_1^M)$ are to be stored for reference, then a very large table is required. With 200 keywords the possible number of combinations of double keywords is 20,100 and there are 1,353,400 combinations of triple keywords, etc. Even though the 1966 acoustic titles do not contain all these possible combinations of double or triple keywords the required tables are still large, and the execution time for table look-up is correspondingly large. There is another problem in that when a request has a new combination of keywords not contained in the tables then no category can be assigned for it. In order to solve these difficulties, Maron assumed that in each category keywords occur statistically independently. He then computed $P(C_k; \{K_i\}_1^M)$ as shown in the formula (4.4) of Chapter IV.

In the present treatment it is supposed that any document can be properly indexed by only one or two keywords. Two tables are therefore

stored in computer memory. One is for the classification of a document indexed by a single keyword, and is called a "single keyword table". The other is for the classification of a document indexed by double keywords, and is called a "double keyword table". In the 1966 acoustic titles, the 200 single keywords produce a total of 560 combinations of double keywords.

The single keyword table contains elements as shown in Table 7.1 for the special case of two keywords and three categories. The columns are formed from the columns of Table 5.1 that correspond to the selected keywords. The categories are those that correspond to the peak values in columns of Table 5.1. Thus the element in the i^{th} row and j^{th} column indicates the number of documents that lie in the i^{th} category and contain the j^{th} keyword. The final row of the table lists the "response category" C_k which contains the most documents associated with the corresponding keyword K_j . If all documents that contain K_j are automatically assigned to the category C_k , then the difference between the number of correct and incorrect assignments is as shown in the "difference" row of Table 7.1.

For example, in Table 7.1, the largest element in the K_1 column is 10. Thus the response category is C_1 , and the difference is $10 - 4 - 3 = 3$.

The double keyword table is similar but the columns correspond to keyword pairs instead of to single keywords. Thus each column of the double keyword table lists the number of documents which contain a given keyword pair, and which are in categories C_1, C_2, \dots, C_{14} . The last but one element of each column of the double keyword table indicates the difference between the number of correct classifications and the number

of incorrect classifications that would result if documents are assigned to the category C_k for which the column element is a maximum. The last element of each column indicates the particular category C_k .

The following steps indicate an algorithm that may be used to classify a document into one of the categories C_k ;

1. Examine the document for the presence of one, or more, of the 200 keywords.
2. If the document has only one keyword, then go to step 3. If it has only two keywords, then go to step 4, otherwise go to step 6. (If no keyword appears on the document, our classification procedure is not applicable.)
3. Look up the single keyword table, and determine the response category. END.
4. Look up the double keyword table. If the keyword pair is in the table, then classify the document under the corresponding response category. END. Otherwise go to step 5.
5. Referring to the difference corresponding to each keyword in the single keyword table, determine which keyword gives the maximum difference. Classify the given document under the corresponding response category. END.
6. Form possible pairs of keywords.
7. If no pair is on the double keyword table, then go to step 5. Otherwise go to step 8.
8. Look up the double keyword table and determine a word pair which has the maximum difference for the possible pairs in the document. Classify the document under the corresponding category. END.

Table 7.1 An Example of Single Keyword Table

(consisted in the instance of only
two keywords and three categories)

	K_1	K_2
C_1	10	1
C_2	4	5
C_3	3	0
Difference	3	4
Response category	C_1	C_2

7.1.2 Experimental Results.

The results of the test are as shown in Table 7.2.

In Table 7.2 the number of titles classified by the single keyword table indicates all titles which contain a single keyword and some titles which contain more than one keyword of which no pair appear on the double keyword table. Examination of part A shows that the accuracy of correct classification by the use of the single keyword table was not satisfactory since it is only 68% for group 1 and 54% for group 2. On the other hand, examination of part B shows that when the double keywords table may be used, the percentages of correct classification are 97.7% for group 1, which is almost perfect, and 89.7% for group 2. This suggests that for this experimental data base it is not necessary to generate a triple keywords table. Part C of Table 7.2 shows the results of the total classification system, which are comparable with Maron's results shown in Table 6.1. For group 1 this method which classified 88.2% of classifiable titles correctly was slightly superior to Maron's one which had 86.3% of correct classifications. On group 2, however, both methods were equally satisfactory.

The most striking difference between Tables 7.2 and 6.1 is in the percentage of correct classifications when the double keyword table may be used. The percentages 97.7, 89.7, 78.8, and 77.8% obtained by use of the double keyword table are significantly higher than the 89.7, 70.6, 65.4, and 64.8 listed in Table 6.1, for documents that contain two, or more, keywords.

To examine the results in more detail we may tabulate them as in Table 7.3 to include the case in which the above steps 1 to 8 rank the response categories in such manner that the correct subject category has

Table 7.2 Experimental Results of Maximization Method.

		group 1 (1966)	group 2 (1967)	group 3 (1968)	group 4 (1961)
Total number of titles	N_t	395	385	506	286
Number of titles with no keyword		82	102	143	96
Number of titles with at least one keyword.	N_k	313	283	363	190
A					
Number of titles clas- sified by the single keyword table.	N_{k1}	100	215	264	136
Number of correct classifications.		68	116	152	67
% of N_{k1}		68.0%	54.0%	57.6%	49.3%
B					
Number of titles clas- sified by the double keyword table.	N_{k2}	213	68	99	54
Number of correct classifications.		208	61	78	42
% of N_{k2}		97.7%	89.7%	78.8%	77.8%
C					
Number of titles classified by either of two tables.	$N_{k1} + N_{k2}$	313	283	363	190
Number of correct classifications.		276	177	230	109
% of $N_{k1} + N_{k2}$ (N_k)		88.2%	62.5%	63.4%	57.4%
% of N_t		69.9%	46.0%	45.5%	38.1%

second rank in the list. In group 1 there are 305 out of 313 titles classified correctly in one of the first two ranks. This is a proportion of 97.4% of the group 1 with at least one keyword. The program following Maron's method printed out 287 as the number of titles having correct categories in one of the first two positions. The proportion was 91.7%. On group 2, our method listed exact categories for 213 titles and Maron's method did for 210 titles out of 283 classifiable titles; the proportions were 75.3% and 74.2% respectively.

It appears that the method of the present section is consistently slightly better than Maron's. One is tempted to conclude that the classification may well be at least as good as would be obtained by manual assignment of categories.

It is envisaged that classification of documents into first and second rank categories might be used in information retrieval as follows. A searcher who requests a list of documents in a certain category would first receive a list of those whose highest ranking response category is that specified. His request could be broadened by addition of the documents whose second highest ranking response is the one specified. Such a broadening would introduce a number of non-relevant items but, according to Table 7.3, would include many relevant documents that were not correctly classified in the highest rank.

The manner in which various documents are classified in the first and second rank is shown in Appendix D. The titles listed are for the documents of group 1. The keywords are underlined.

One of the most significant features of Table 7.3 is its indication of how close Maron's method comes to approaching the accuracy of the method that does not depend on the assumption of statistical independence.

Table 7.3 Comparison of Maximization Method with Maron's Method

	Maximization Method	Maron's Method
Group 1 (1966)		
Number of titles with at least one keyword, N_k	313	313
Number of correct classifications listed in the first rank.	276	270
% of N_k	88.2%	86.3%
Number of correct classifications listed in the second rank.	29	17
Number of correct classifications listed in one of the first two ranks.	305	287
% of N_k	97.4%	91.7%
Group 2 (1967)		
Number of titles with at least one keyword, N_k	283	283
Number of correct classifications listed in the first rank.	177	176
% of N_k	62.5%	62.2%
Number of correct classifications listed in the second rank.	36	34
Number of correct classifications listed in one of the first two ranks.	213	210
% of N_k	75.3%	74.2%
Group 3 (1968)		
Number of titles with at least one keyword, N_k	363	363
Number of correct classifications listed in the first rank.	231	216
% of N_k	63.6%	59.5%
Number of correct classifications listed in the second rank.	44	45
Number of correct classifications listed in one of the first two ranks.	275	261
% of N_k	75.8%	71.9%
Group 4 (1961)		
Number of titles with at least one keyword, N_k	190	190
Number of correct classifications listed in the first rank.	109	108
% of N_k	57.4%	56.8%
Number of correct classifications listed in the second rank.	29	24
Number of correct classifications listed in one of the first two ranks.	138	132
% of N_k	72.6%	69.5%

However, it should be remarked that this is also a verification of our method of choice of keywords. Maron did not use our method, and with alternative choice of keywords the results of Maron's method might be much poorer.

It should also be borne in mind that Maron's procedure is based on an approximation to the joint probability of occurrence of all the keywords in the document. In contrast, our method is based on a knowledge of the exact frequencies of occurrence of single, and pairs of, words. It appears, therefore, that exact information about occurrences of pairs of keywords is somewhat more useful than approximate predictions of frequencies of higher numbers of occurrences.

7.1.3 Suggestions and Discussion.

The hypothesis of the method described in the previous sections is that any document can be indexed by use of tables based on statistical distributions of single or pairs of keywords. In accordance with this hypothesis the single keyword table and double keyword table were prepared for the acoustic titles.

As was mentioned in section 7.1.1, when a document to be indexed has a pair of keywords K_1 and K_2 , but the double keyword table does not include such a pair, then one of K_1 and K_2 is chosen according to which corresponds to the larger difference value in the single keyword table. Thus the existence of one keyword in the document is completely neglected. One suggestion may be made to avoid this irrationality.

Instead of choosing one keyword from the pair K_1 and K_2 , evaluate the number of documents classified by K_1 or K_2 , for example in Table 7.1 take the sum of the two column vectors $K_1 = \begin{pmatrix} 10 \\ 4 \\ 3 \end{pmatrix}$ and $K_2 = \begin{pmatrix} 1 \\ 5 \\ 0 \end{pmatrix}$ to get

$K_1 + K_2 = \begin{pmatrix} 11 \\ 9 \\ 3 \end{pmatrix}$. The category C_1 in which the largest number of doc-

uments can be classified correctly by either K_1 or K_2 may be chosen as the response category for the request. By this improvement, all the keywords in a request can participate in determining the response category without neglecting any of the keywords.

7.2 Modification of Maron's Method Using Keyword Association.

7.2.1 Classification System Based on Keyword Association.

By knowing the relationships between keywords, a document can be extended by some additional keywords, and then the classification system will be able to assign more correctly the category to the document.

A measure of the degree in which keywords are associated within documents may be formulated as follows. Suppose that $N(K_i)$ and $N(K_j)$ are the frequencies with which documents are indexed by keywords K_i and K_j respectively. Let $N(K_i, K_j)$ be the frequency with which both K_i and K_j index documents. The probability that a document containing K_i also contains K_j may be computed as

$$P(K_j; K_i) = \frac{N(K_i, K_j)}{N(K_i)} \quad (7.1)$$

which gives a measure of the extent to which K_j tends to occur in documents that contain K_i . If no document contains both K_i and K_j , then $P(K_j; K_i) = 0$. If every document containing K_i also contains K_j , then $P(K_j; K_i) = 1$. In general, $P(K_j; K_i) \neq P(K_i; K_j)$, since for example, the word "retrieval" tends to be used with the word "information", but "information" is often used without association with "retrieval".

The proposed application of a keyword association technique results in a modification of Maron's classification method as described below. Assume that a request document is indexed by only one keyword, indicated by K_i . The keyword K_i is associated with the keyword K_r ($r=1$ to 200) to an extent measured by $P(K_r;K_i)$. Obviously the degree of association of K_i with itself is equal to 1, viz. $P(K_i;K_i) = 1$. The probability of $P(C_k;K_i)$ that a document indexed by K_i belongs to category C_k is modified as follows:

$$P(C_k;K_i) \approx \sum_{r=1}^{200} P(C_k;K_r)P(K_r;K_i) \quad (7.2)$$

In formula (7.2) each term that appears on the right-hand side denotes the individual attribute number computed between category C_k and keyword K_r which relates to the given keyword K_i by a degree $P(K_r;K_i)$. If a keyword K_r is closely related with K_i then the attribute number $P(C_k;K_r)$ is considered as an important factor. Thus the probability $P(K_r;K_i)$ may be regarded as the weight through which $P(C_k;K_r)$ contributes to the value of $P(C_k;K_i)$.

Next, assume that a given document is indexed by M number of keywords, indicated by $\{K_i\}_1^M$. Maron derived the formula of an attribute number as follows:

$$P(C_k;\{K_i\}_1^M) \approx P(C_k) \prod_{i=1}^M P(K_i;C_k) \quad (7.3)$$

which is given in formula (4.4) of Chapter IV. There is a statistical relation between probabilities as follows:

$$P(K_i; C_k) = \frac{P(C_k; K_i)P(K_i)}{P(C_k)} \quad (7.4)$$

Substitute (7.4) into (7.3) that

$$P(C_k; \{K_i\}_1^M) \approx P(C_k) \prod_i^M \frac{P(C_k; K_i)P(K_i)}{P(C_k)} \quad (7.5)$$

where $\prod_i^M P(K_i)$ is independent of categories and therefore its computation is unnecessary. The formula (7.5) may therefore be simplified to the form

$$P(C_k; \{K_i\}_1^M) \approx P(C_k)^{1-M} \prod_i^M P(C_k; K_i) \quad (7.6)$$

Substituting the formula (7.2) into (7.6) then gives

$$P(C_k; \{K_i\}_1^M) \approx P(C_k)^{1-M} \prod_{i=1}^M \left\{ \sum_{r=1}^{200} P(C_k; K_r)P(K_r; K_i) \right\} \quad (7.7)$$

where k varies from 1 to 14.

The derived formula (7.7) is the general form of attribute number modified by keyword association. We call the resulting number a "modified attribute number". The necessary probabilities $P(C_k)$, $P(C_k; K_r)$ and $P(K_r; K_i)$ for the computation of the modified attribute number are defined in the following manner;

$$P(C_k) = \frac{\text{number of documents belonging to the } k^{\text{th}} \text{ category}}{\text{total number of documents}}$$

$$P(C_k; K_r) = \frac{\text{number of documents with the } r^{\text{th}} \text{ keyword belonging to the } k^{\text{th}} \text{ category}}{\text{number of documents containing the } r^{\text{th}} \text{ keyword}}$$

$$P(K_r; K_i) = \frac{\text{number of documents containing both the } i^{\text{th}} \text{ and } r^{\text{th}} \text{ keywords}}{\text{number of documents containing the } i^{\text{th}} \text{ keyword}}$$

7.2.2 Experimental Results.

Following the manner of the previous experiments, the modified classification system was tested on acoustic 1966, 1967, 1968, and 1961 data separately. The figures derived from this experiment are shown in Table 7.4.

In Table 7.4, the results of Maron's method from Table 7.3 are , repeated in order to clarify the comparison with our modified method.

It is clear that the modified method has no advantage over Maron's method. In fact it is slightly poorer than the previous method whose results are in Table 7.3. This suggests that information about word associations is less useful than information about the words actually present in the documents.

The above fact is not surprising in view of the extremely careful way of choosing the keywords. If the keywords had been chosen in a less optimum manner, then some important keywords K_r might have been omitted, in which case they could influence the choice of category only through the effect of non-zero values of $P(K_r; K_i)$.

7.2.3 Discussion.

Comparing two methods from the results in Table 7.4 for the titles classified correctly in the first rank, the two methods were very similar in their determination of categories. But for group 2, the modified method performed poorly. In group 1, the modified method

Table 7.4 Experimental Results of Modified Method and Their Comparison with Those of Maron's Method

	Modified Method	Maron's Method
Group 1 (1966)		
Number of titles with at least one keyword, N_k	313	313
Number of correct classifications listed in the first rank.	270	270
% of N_k	86.3%	86.3%
Number of correct classifications listed in the second rank.	24	17
Number of correct classifications listed in one of the first two ranks.	294	287
% of N_k	93.9%	91.7%
Group 2 (1967)		
Number of titles with at least one keyword, N_k	283	283
Number of correct classifications listed in the first rank.	175	176
% of N_k	61.8%	62.2%
Number of correct classifications listed in the second rank.	26	34
Number of correct classifications listed in one of the first two ranks.	201	210
% of N_k	71.0%	74.2%
Group 3 (1968)		
Number of titles with at least one keyword, N_k	363	363
Number of correct classifications listed in the first rank.	207	216
% of N_k	57.0%	59.5%
Number of correct classifications listed in the second rank.	53	45
Number of correct classifications listed in one of the first two ranks.	260	261
% of N_k	71.6%	71.9%
Group 4 (1961)		
Number of titles with at least one keyword, N_k	190	190
Number of correct classifications listed in the first rank.	99	108
% of N_k	52.1%	56.8%
Number of correct classifications listed in the second rank.	28	24
Number of correct classifications listed in one of the first two ranks.	127	132
% of N_k	66.8%	69.5%

produced some improvement in choice of correct classifications for the titles listed in the second rank.

The above facts may imply the following conclusions. For the titles which Maron's method classified correctly in the first rank, keywords of these titles are strongly associated with their correct categories. Therefore the values of $P(K_r; K_i)$ used in the modified method hardly affect the choice of correct categories for such titles. On the other hand the titles classified correctly in the second rank may be regarded as having relatively weak associations with their correct categories, in which case the attribute numbers $P(C_k; K_r)P(K_r; K_i)$ of such keywords K_r that are strongly associated with keywords K_i in a title give rise to the better classification.

From Appendix B, which indicates the similarity coefficients between three groups of data, it may be seen that the similarity coefficient between group 1 and group 2 has the lowest value. This implies that the behavior of the keyword distributions of group 1 differs somewhat from that of group 2. This may cause a decrease in the number of correct classifications for the second rank in group 2.

From the results shown in Table 7.4 it appears that the attribute numbers provide a useful means to classify documents into categories, and that a great deal of improvement cannot be expected by the use of the modified method.

7.2.4 Suggestion.

In the modified method it is assumed that every keyword originally appearing on a document is equally significant. This assumption may not be realistic. When we analyze a document, first we look for the

important sentences and words, and next we rank them in significant order.

A further suggestion is to generate a classification system that involves the concept that every keyword in a document has a significant factor, or weight. Suppose that a request document contains M keywords, indicated by $\{K_i\}_1^M$ including additional keywords, and that by some method all weights of these keywords, indicated by $\{w_i\}_1^M$, are determined and that the weight w_i is independent of categories. $P(C_k; \{w_i K_i\}_1^M)$ defines the probability that the request document indexed by $\{K_i\}_1^M$ with weights $\{w_i\}_1^M$ belongs to category C_k .

Let us make an assumption that the weight of a keyword is also the weight of the probability that a document in a category C_k may be indexed by this keyword. This can be formulated as

$$P(w_i K_i; C_k) = w_i P(K_i; C_k) \quad (7.8)$$

For the approximation of the value $P(C_k; \{w_i K_i\}_1^M)$ two possible approaches may be used. First, the value may be set as the sum of the attribute numbers between each keyword K_i and category C_k multiplied by the weight w_i ; thus;

$$P(C_k; \{w_i K_i\}_1^M) \approx \sum_i^M w_i P(C_k; K_i) \quad (7.9)$$

The form of the right hand side of (7.9) ensures that if a document contains many keywords of high weights strongly associated with a certain category, then the attribute number of the document becomes large. Even if a keyword is strongly associated with a category, but the weight of

the keyword in the document is small, then the term on the right hand side of (7.9) cannot be considered to be important.

Another approach is based on the assumption that in each category the keywords occur statistically independently. The attribute number $P(C_k; \{w_i, K_i\}_1^M)$ may then be modified to become

$$P(C_k; \{w_i, K_i\}_1^M) = \frac{P(\{w_i, K_i\}_1^M; C_k) P(C_k)}{P(\{w_i, K_i\}_1^M)} \quad (7.10)$$

where the denominator $P(\{w_i, K_i\}_1^M)$ is independent of categories and hence may be eliminated. By the assumption of keyword independence, the formula (7.10) is simplified as follows:

$$P(C_k; \{w_i, K_i\}_1^M) \approx P(C_k) \prod_i^M P(w_i, K_i; C_k) \quad (7.11)$$

Substitution of the relations (7.8) into (7.11) leads to the formula

$$P(C_k; \{w_i, K_i\}_1^M) \approx P(C_k) \prod_i^M w_i P(K_i; C_k) \quad (7.12)$$

The method proposed by Stiles in 1961 (20) appears suitable for estimation of the value of the weights. It not only computes the weights of keywords, but also produces the additional keywords to extend a request document as well. The precise procedure is shown in Appendix C.

We have not tested the classification methods suggested in the formulae (7.9) and (7.12).

CHAPTER VIII

CONCLUSIONS

The present thesis has studied automatic classification systems based on statistical relationships between words and subject categories of documents. Several experimental trials have been described.

Using the IBM 360/67 computer installed at The University of Alberta Computing Center, experiments were designed using titles and authors published by JASA (the Journal of the Acoustical Society of America) in 1966, in 1967, in 1968, and in 1961.

Chapter II and Chapter III contain an examination of the applicability of latent class analysis to document classification. The latent class analysis is critically dependent on the assumption that, in each latent class, keywords occur statistically independently. This assumption is expressed through the form of the accounting equations (2.5), (2.6) and (2.7) in Chapter II. From the analysis of word occurrences latent class analysis can provide a set of classification categories as well as probabilities between words and categories. The examination of latent class analysis provided clear illustration of its unsuitability for document classification systems. Determination of the number of latent classes is a very difficult problem. The hypothesis of Winters that the number of latent classes is equal to the number of keywords facilitates the numerical solution of the accounting equations, but our attempt to apply Winters' technique was unsuccessful. Moreover, it is doubtful whether the latent classes derived from the latent class analysis are meaningful. It must be concluded that the strictly mathematical latent class analysis is not a useful tool with which to

attack the problem of document classification.

Essentially, the same assumptions are used to justify Maron's attribute analysis. Both analyses assume the statistical independence of keyword occurrence in each subject category or latent class. The difference, however, exists in the procedure used to determine the classification schedule. Attribute analysis requires a set of base data which are classified correctly according to a pre-existing classification schedule. In contrast the latent class analysis uses neither pre-existing base data nor a classification schedule. In Chapter IV it was shown that use of attribute numbers for documents keyworded by more than one word assigns a correct category very successfully. It was concluded that attribute analysis forms a promising method for document classification.

One of the methods proposed in the present thesis is a maximization method of correct classification as described in Section 7.1 of Chapter VII. For the documents keyworded by single, or pair of, words the maximization method uses direct statistical descriptions of the base data instead of approximations as calculated in Maron's method based on the independence assumption. In comparison with Maron's method, the experimental results appear to be slightly improved. The results suggest that Maron's approximation that keywords occur statistically independently in each subject category holds meaningfully among documents whose keywords are chosen from natural language.

A modification of Maron's method in terms of keyword associations was proposed in an attempt to improve the classification of documents that contain relatively few keywords. However, the method did not lead to improved classification. It appears that the keywords that are

themselves contained in a document are better clues for assignment of correct categories than are extended keywords derived from keyword associations.

The above fact may imply that the scheme used to select 200 keywords for the present experiments helps a great deal to ensure correct classifications. If the keywords are selected less carefully, however, the attribute numbers corresponding to extended keywords, $P(C_k; K_r)P(K_r; K_i)$, in the modified method may be necessary in order to extend a request document and its correct category.

Throughout the experiments it was found that the proposed method of choice of keywords is very suitable for the classification of document titles. Titles used in the present experiments contain an average of 8 or 9 words and 1 or 2 keywords. Therefore, the use of direct statistics on occurrences of more than two keywords was impossible. However, the direct application of this method to the classification of abstracts or full text may involve some problems relating to the memory size and execution time of the computer.

One of the conclusions of the present investigation is that document titles may provide a very useful source of keywords for classification. For the acoustics data base the classification effectiveness does not significantly change with respect to time except for an initial reduction when the data is extended beyond the base documents.

REFERENCES

1. Anderson, T. W., "On Estimation of Parameters in Latent Structure Analysis", *Psychometrika*, Vol. 19, No. 1, pp. 1 - 10, March, 1954.
2. Baker, F. B., "Information Retrieval Based upon Latent Class Analysis", *Journal of the ACM*, Vol. 9, No. 4, pp. 512 - 521, September, 1962.
3. Baker, F. B., "Latent Class Analysis as an Association Model for Information Retrieval", In: *Symposium on Statistical Association Methods for Mechanized Documentation*, Washington, D.C., 1964. *Proceedings*, Washington, D.C., National Bureau of Standards, 1964. *Miscellaneous Publications* 269, pp. 149 - 156.
4. Decimal Classification (D.C.), Batty, C. D., "An Introduction to The Dewey Decimal Classification", London, C. Bingley, 1965.
5. Foskett, D. J., "The Construction of a Faceted Classification for a Special Subject", In: *Proceedings of International Conference on Scientific Information*, Vol. 2, pp. 868 - 888, 1958.
6. Green, B. F., Jr., "A General Solution for the Latent Class Model of Latent Structure Analysis", *Psychometrika*, Vol. 16, pp. 151 - 166.
7. Heaps, D. M. and Heaps, H. S., "Computer Retrieval of Papers in Acoustics", *Journal of the Acoustical Society of America*, Vol. 43, No. 6, pp. 1461 - 1463, June, 1968.
8. Lazarsfeld, P. F. and Henry, N. W., "Latent Structure Analysis", Houghton Mifflin Company, 1968.

9. Library of Congress Classification (L.C.) "Institute on The Use of Library of Congress Classification", New York, 1966.
10. Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Library Information", IBM Journal of Research and Development, Vol. 1, pp. 309 - 317, 1957.
11. Luhn, H. P., "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, Vol. 2, pp. 159 - 165, 1958.
12. Maron, M. E. and Kuhns, J. L., "On Relevance, Probabilistic Indexing and Information Retrieval", Journal of the ACM, Vol. 7, pp. 216 - 244, 1960.
13. Maron, M. E., "Automatic Indexing: An Experimental Inquiry", Journal of the ACM, Vol. 8, No. 3, pp. 404 - 417, 1961.
14. Ranganathan, Shiyali Ramamrita, rao sahib, "Colon Classification", Bombay Asia Publishing House, 1963.
15. Ranganathan, Shiyali Ramamrita, rao sahib, "Documentation and Its Facets; Being A Symposium of Seventy Papers by Thirty-Two Authors", Bombay, New York, Asia Publishing House, 1963.
16. Richardson, E. C., "Classification", The 3rd Edition, The Shoe String Press, Inc., 1964.
17. Salton, G., "Document Retrieval System for Man-Machine Interaction", The ACM Proceedings of the 19th National Conference, 1964.
18. Salton, G., "Automatic Information Organization and Retrieval", McGraw-Hill Series in Computer Science, 1968.
19. Shannon, C. E., "The Mathematical Theory of Communication", Urbana, University of Illinois Press, 1964.

20. Stiles, H. E., "The Association Factor in Information Retrieval",
Journal of the ACM, Vol. 8, No. 2, pp. 271 - 279, 1961.
21. Todd, John, "Survey of Numerical Analysis", New York, McGraw-Hill,
1962.
22. Universal Decimal Classification (U.D.C.) Freeman, Robert R.,
"File Organization and Search Strategy Using Reference
Retrieval Systems", New York, America Institute of Physics.
23. Vickery, B. C., "Faceted Classification, A Guide to Construction
and Use of Special Schemes", Prepared for The Classification
Research Group, London, Aslib, 1960.
24. Winters, W. K., "A Modified Method of Latent Class Analysis for
File Organization in Information Retrieval", Journal of the
ACM, Vol. 12, No. 3, pp. 356 - 363, July, 1965.

APPENDIX

APPENDIX A: LIST OF PARTICIPANTS

Participant ID	Age	Gender	Education	Occupation	Marital Status
001	25	Male	High School	Student	Single
002	28	Female	College	Teacher	Married
003	30	Male	University	Engineer	Single
004	32	Female	High School	Homemaker	Married
005	35	Male	College	Manager	Married
006	38	Female	University	Doctor	Single
007	40	Male	High School	Worker	Married
008	42	Female	College	Librarian	Married
009	45	Male	University	Professor	Single
010	48	Female	High School	Retired	Married
011	50	Male	College	Business	Married
012	52	Female	University	Researcher	Single
013	55	Male	High School	Farmer	Married
014	58	Female	College	Nurse	Married
015	60	Male	University	Lawyer	Single
016	62	Female	High School	Homemaker	Married
017	65	Male	College	Retired	Married
018	68	Female	University	Professor	Single
019	70	Male	High School	Worker	Married
020	72	Female	College	Librarian	Married
021	75	Male	University	Retired	Married
022	78	Female	High School	Homemaker	Married
023	80	Male	College	Retired	Married
024	82	Female	University	Retired	Married
025	85	Male	High School	Retired	Married
026	88	Female	College	Retired	Married
027	90	Male	University	Retired	Married
028	92	Female	High School	Retired	Married
029	95	Male	College	Retired	Married
030	98	Female	University	Retired	Married

APPENDIX A

List of Keywords Chosen as Described in Chapter V

1	ABSEN	21	BOER	41	CLICK	61	DUE	81	FLORI
2	AMBIE	22	BOOMS	42	COCHL	62	DURAT	82	FLOW
3	APPRO	23	BOOM	43	CONIC	63	DURLA	83	FREE
4	ARASE	24	BOTTO	44	CONSO	64	EAR	84	F2
5	ARCTI	25	BROWN	45	CRYST	65	EARCA	85	GELLE
6	ATTAC	26	BULLF	46	CYLIN	66	EARPH	86	GEOME
7	AUDIO	27	BURKE	47	DALLO	67	EC	87	GOLD
8	AUDIT	28	BURST	48	DAMPE	68	ELEME	88	GOODM
9	AXIAL	29	CABLE	49	DAMPI	69	ELFNE	89	GOULD
10	AXISY	30	CALIB	50	DATA	70	ENGLI	90	GUINE
11	BATCH	31	CALLA	51	DEATH	71	ERRAT	91	HAVIN
12	BAUER	32	CAMPB	52	DECIS	72	EXAMI	92	HEARI
13	BEAMS	33	CANTI	53	DEEP	73	EXIST	93	HECKE
14	BEATT	34	CARHA	54	DENHA	74	EXPLO	94	HENNI
15	BENZE	35	CAROM	55	DICHO	75	EXPOS	95	HODGE
16	BIBLI	36	CARTE	56	DIRKS	76	FAR	96	HOLLO
17	BILGE	37	CAUSE	57	DISCR	77	FATIG	97	HUMAN
18	BINAU	38	CHANN	58	DITAR	78	FILTE	98	HUTTO
19	CLASI	39	CLACK	59	DOBBI	79	FITZG	99	HYDRO
20	BOBBE	40	CLAMP	60	DOOLI	80	FLEXU	100	HYPER

101	ICE	121	LOUDN	141	OXYGE	161	RODS	181	STRIK
102	IMPAC	122	MASKE	142	PANEL	162	ROOMS	182	SUBHA
103	INDUC	123	MASKI	143	PEOPL	163	SACKM	183	TANG
104	INTEL	124	MASSE	144	PERIP	164	SANDW	184	THIN
105	INTEN	125	MCCOM	145	PIG	165	SEA	185	THREE
106	INTRA	126	MCNIV	146	PISTO	166	SECTI	186	THRES
107	JOHNS	127	MELLE	147	PITCH	167	SEGME	187	TILLM
108	JOURN	128	METAL	148	POTEN	168	SEGUI	188	TONAL
109	KARNO	129	MIGHT	149	POWER	169	SHAH	189	TONES
110	KC	130	MODEL	150	PROGR	170	SHALL	190	TORIC
111	LAMB	131	MOVIN	151	RADIU	171	SHAW	191	TUBE
112	LAMIN	132	MUSCL	152	RAYS	172	SHELL	192	UBERA
113	LATER	133	NERVE	153	RECIP	173	SHIFT	193	ULTRA
114	LAW	134	NEURA	154	REFER	174	SIGNA	194	UNDER
115	LAYER	135	OCEAN	155	RELAX	175	SINGH	195	UNMAS
116	LINDS	136	OHMAN	156	REMOT	176	SINUS	196	VIBRA
117	LINNE	137	OLSEN	157	REVER	177	SONIC	197	VOCOD
118	LIQUI	138	ORIGI	158	RIGID	178	SOURC	198	WEST
119	LOAD	139	ORTHO	159	RING	179	SPEEC	199	WHY
120	LONGI	140	OUT	160	ROD	180	STORY	200	ZERLI

APPENDIX B

Similarity of Successive Years of Data Base

It has been supposed that the statistics of the base data are similar to those of the data to be classified.

Suppose that a group of data is represented in vector form where each element of a vector indicates the frequency of occurrence of the corresponding word on the data, and that there are two groups of data such as;

$$T = (t_i)_{i=1,N} \quad (B-1)$$

$$U = (u_i)_{i=1,N}$$

where N is the total number of different words on two groups.

The cosine coefficient is computed as follows:

$$\frac{\sum_{i=1}^N t_i u_i}{\sqrt{\sum_{i=1}^N t_i^2 \sum_{i=1}^N u_i^2}} = \text{cosine coefficient} \quad (B-2)$$

In N dimensional space, if the angle between T and U is 0° then the cosine coefficient is 1, which means that the two groups of data are identical except for an amplitude factor. If the angle is 90° , then the cosine coefficient is 0, which means that there are no common keywords that index the two groups of data.

The formula (B-2) has been used to compute the similarity between the acoustic 1966, 1967, and 1968 titles. The results are shown in

matrix form as follows:

	(1966)	(1967)	(1968)	
(1966)	1.0	0.85709	0.86216	(B-3)
(1967)	0.85709	1.0	0.87804	
(1968)	0.86216	0.87804	1.0	

Among the coefficients in (B-3) the one between acoustic 1967 and acoustic 1968 titles has the highest value, the next highest one is between acoustic 1966 and acoustic 1968 titles. Therefore, the choice of acoustic 1968 as a base data is the best among three groups. However the correlation is not significantly different for any two of the three years.

Stiles' Measure of Association Factor and Choice of Keywords

In his paper (20) Stiles introduced a formula to measure the degree of association between two keywords and named it an "association factor" defined as follows:

$$\log_{10} \frac{(|N_{ab}N - N_aN_b| - \frac{N}{2})^2 N}{N_aN_b(N - N_a)(N - N_b)} = \text{association factor} \quad (C-1)$$

where N is the total number of documents; N_a is the number of documents indexed by word A ; N_b is the number of documents indexed by word B ; and N_{ab} is the number of documents indexed by both A and B . If $N_{ab}N$ is less than N_aN_b , the association factor must be considered negative. In the computation of the association factor between word A and itself, $N_{aa} = N_a$.

The following steps describe how to generate the additional keywords to extend a request document and how to determine the weights of the keywords.

1. For each keyword on a request document, form a "profile" consisting of all keywords which, in association with the given one, have association factors greater than 1.0.
2. From the profiles of all keywords on the document, select additional keywords that appear frequently in the set of profiles. They are called the first generation keywords.
3. For the original keywords and the first generation keywords, repeat step 1 and step 2 and select second generation keywords.
4. For each keyword, including the original keywords, the first generation keywords and the second generation keywords, compute the

sum of the association factors on its profile. The sum is called the weight of the keyword. This weight is a measurement of the degree of association between the keyword and the request document.

APPENDIX D

First and Second Rank Classification

Using Modified Attribute Analysis

pp. 89 - 97.

(The notation of 14 categories A, B, ...,N used on the following list corresponds to that of the categories 1, 2, ..., 14 used in section 5.2, respectively. Keywords are underlined.)

	CORRECT CATEGORY	ASSIGNED CATEGORIES	
		1ST RANK	2ND RANK
AC066039I	1DYN	KLCN AVISY PLANE STRAI DYNAM RESPO THICK ORTHO SHELL	I, H,
AC066039I	8EUGI	PENGE FREE VIBRA THIN ISOTR OBLAT SPHER SHELL	I, H,
AC066039I	14YEH	FORCE VIBRA TAO DEGRE FREED SYSTE COMBI COULO VISCO DAMPL	I, G, H,
AC066039J	25TCPIC	BAUER EXPR STUDI UNDER EARP H LOUDN BALAN METHO	J, B,
AC066039J	35TOPIC	BAUER CALLB ANALY UNDER EARP H LOUDN BALAN METHO	J, B,
AC066039J	40EAT	RECIP CALLIE IDEE ACTIV IMEED TERM	J, H,
AC066039J	48PHILL	BOERE LEATI SONAR TRANS CALLIE HIGH PRESS TUBE	J, H,
AC066039J	55	NOI CLASSIFIABLE	J,
AC066039L	62PRAND	SIENW ALLEN PSEUD SIGNA CORRE METHO UNDER ACOUS RESEA	L,
AC066039L	II	INSTR	L,
AC066039L	74SHOR	ADAPT TECHN DISCR AGAIN COHER NOISE NARRO BAND SYSTE	L,
AC066039L	79USHER	SIGNA DETEC ARRAY ARBIT PFOCE DETEC	L,
AC066039B	87SUTTO	BABKO END POINT LATER DICHQ CLICK	B,
AC066039B	103BILGE	BEFOI MARKI ABSEN INTRA AURAL MUSCL	B, F, J,
AC066039B	109ZERLI	DAVIS ACOUS RELAT HUMAN VERTE POTEN	M, H,
AC066039B	117ELLIO	SILBI AUDIT THRES LCCAT UNCER FUNCT TONE PARAM FATIG	D, J,
AC066039B	125CLSEN	JOHNS TILLM EARP H VERSU SOUND FIELD THRES SOUND PRESS IEVEL	C, J,
AC066039B	134ZERLI	INTER TIME INTEN DIFFR MLD	G, J,
AC066039F	138WILLI	KRYTE MASKI SPECFC AIRCR NOISE	B, F,
AC066039F	151CHMAN	COABT VCV UTTER SPECT	F,
AC066039G	169	NOT CLASSIFIABLE	M,
AC066039G	170CRUM	STUMP ULTRA STUDI MOLEC ASSOC AQUEO SOLUT FORMI ACETI PROPI	G,
AC066039H	171	NOT CLASSIFIABLE	I,
AC066039J	173GOODH	GRACE CIRCU WAVES SOLID CYLIN	J,
AC066039J	231LINDS	NEW DEVEL JCUN	B,
AC066039B	232DEATH	EXAMI BINAU INTER	G,
AC066039G	250FJII	YAMAD ULTRA ATTEN RELAX TIMES WATER VAPOR HEAVY WATER VAPOR	H,
AC066039H	255GUPTA	REFIE SCUND WAVES TRANS LAYER	I,
AC066039I	261EOGDA	UNICU MEAN SQUAR APPRO SYSTE	I,
AC066039I	269FRANK	RAYNO METHC COMPU LAMPE RESCN FREQU SYSTE EQUAL MASSE EQUAL	I,
AC066039J	272UBERA	DOCLI SCUND SCATT ELAST CYLIN SHELL	I,
AC066039J	276HICKL	WANG SCATT SOUND RIGID MCVAB SPHER	I,
AC066039J	280LIGGE	JACCB NCISE CCVAR VERTI LIRIC DEEP OCEAN	J,
AC066039J	289	NOT CLASSIFIABLE	J,
AC066039J	294RUDGE	MCNOS REFLE RIGID OBJEC DEFIN QUADR SURPA	J,
AC066039J	301STEIN	BIRDS UNDER SOUND PROP	J,
AC066039J	316	NOT CLASSIFIABLE	J,
AC066039L	323	NOT CLASSIFIABLE	J,
AC066039B	336HENNI	FREQU LISCR RANDO ANPII TCNES	J,
AC066039B	340PPAFF	MATHE LETEC AUDIT SIGMA REPRC NOISE	J,
AC066039B	346ILCYD	REDWO FINIT DIFFE METHC INVES VIBRA SOLID EVALU EQUIV CIRCU	J,
AC066039D	362TOVE	CHARA PIEZO RESCN	J,
AC066039D	372SINGH	SCNIC WAVEP	J,
AC066039F	388WICKE	DISII FEATU ERROR SHORT TERM MEMOR ENGLI CONSO	J,

	CORRECT CATEGORY	ASSIGNED CATEGORIES	
		1ST RANK	2ND RANK
AC066039B 720GREEN INTER PHASE EFFEC MASKI SIGNA DIFFE DURAT	B	B,	F,
AC066039B 725STEVE PCWER GROUF TRANS UNDER GLARE MASKI RECRU	B	B,	J,
AC066039B 736WARD USE SENSE LEVEL MEASU ICUDN TEMPO THRES SHIFT	B	B,	D,
AC066039C 741			
AC066039B 748GREEN COMME EFFEC WAVEF CORRE SIGNA DURAT DETEC NOISE BURST CONTI	B	B,	F,
AC066039C 749			
AC066039G 751CHIAO FLEUR DISPE HYPER WAVES LIQUI	G	G,	I,
AC066039G 752SEGUI LAMB SURFA TREAT QUART CRYST HYPER APPLI	G	G,	I,
AC066039H 753			
AC066039I 755NEWLA CCMME VIERA ENERG TRANS THREE ELEME STRUC	I	I,	J,L,
AC066039I 755SSCHAR LYCN REPLY CCMME VIBRA ENERG TRANS THREE	I	I,	J,L,
AC066039G 813HERZF FIFTY YEARS PHYSI ULTRA	G	G,	M,
AC066039H 826EURRE LOW FREQU SCATT SOFT SPHER	H	H,	
AC066039H 832			
AC066039H 841			
AC066039I 847WRIGH CHEN FREQU EQUAT WAVE PROP A INITI STRES	I	I,	H,J,K,
AC066039I 849			
AC066039I 856FITZG PARTI WAVES AUDIO MODES CRYST	I	I,	B,
AC066039I 870FITZG OESER NONEL AUDIO RESON	I	I,	B,
AC066039I 878SNCWD VIERA CANTI BEAMS DYNAM	I	I,	G,H,
AC066039I 887UNGAR STEAD STATE RESPO CNE	I	I,	C,
AC066039I 895WILKI TRANS RESPO THIN ELAST SHELL	I	I,	H,
AC066039J 899			
AC066039J 904URICK LONG RANGE DEEP SFA ATTEN MEASU	J	J,	
AC066039A 907BISHO REDUC AIRCR NOISE MEASU SEVER SCHOC MOTEL RESID ROOMS	A	A,	
AC066039B 914			
AC066039B 920ROSEN PITCH DISCR JITTE PULSE TRAIN	B	B,	F,
AC066039C 929HUNT SPENC COUPL THICK SHEAR FLEXU DISPL RECTA QUART PLATE	C	I,	C,
AC066039D 936HUNDI BACKU WALL VIEFA FLUE ORGAN PIPES THEIR EFFEC TONE	D	I,	G,H,
AC066039F 946WILLI HECKE CHCIC REFER CONDI SEEEC PREFE TESTS	F	F,	
AC066039F 953COULT PICKER STALL F2 ADJAC CCNSO PREDI F2 ONSET	F	F,	D,
AC066039G 960			
AC066039G 965			
AC066039G 972PROCT LOW TEMEE SPEED SCUND SINGL CRYST ICE	G	G,	J,J,
AC066039B 978BILGE REMOT MASKI ABSEN INTRA AURAL MUSCL	B	B,	F,J,
AC066039C 978			
AC066039 978WARD WHY STRIK OUT MIGHT KC	I	I,	G,H,
AC066039I 979ANDRE DEFER LASER INTER TECHN MEASU SMALL ORDER VIBRA DISPL	K	K,	
AC066039K 980KFED AMELI VARIA EXPLO WAVES LONG RANGE	E	E,	D,
AC066039E S1HURBA NATUR SONIC BOOM BOOM PROBI	E	E,	D,
AC066039E S10MORRI MACK CARLS SONIC BOOM	E	E,	D,
AC066039E S19SHROU MCLEA DESIG METEC MINIM SCNIC BOOM	E	E,	D,
AC066039E S26KANE SOME EFFEC NONCN ATMCS PROP A SONIC BOOMS	E	E,	D,
AC066039E S31NEWMA HILTO INSTR TECHN MEASU SCNIC BOOM	E	E,	B,
AC066039E S36MAGLI SOME EFFEC AIRFL OPERA ATMCS SONIC	E	E,	B,
AC066039E S43GIFERK EFFEC SONIC BOOM	E	E,	D,
AC066039E S51BORSK NIXCN EFFEC SCNIC BOOM	E	E,	D,
AC066039E S59WARRE EXPER UNITE KINGD EFFEC SCNIC BANGS	E	E,	D,
AC066039E S65KRYTE LABCR TESIS PHYSI PSYCH REACT SONIC BOOMS	E	E,	D,

	CORRECT CATEGORY	ASSIGNED CATEGORIES	
		1ST RANK	2ND RANK
AC066039E S73FOSS	PROBL FUTUR AIR	TRANS VEHIC	
AC066039K1019	NOT CLASSIFIABLE		D,
AC066039K1027SMITH	PURE TONE	FIELD	
AC066039B1030CAMBE	CODIN		G,J, D,J, F, L,
AC066039B1034STUCK	THRES	PHENO	
AC066039B1037JOHNS	SPEFC	PERIO MODUL	
AC066039B1051WYMAN	VARIA	SIGNA DETEC	
AC066039B1056FINCK	RESPO	EQUAL DEVIC	F,J, M,
AC066039B1063KORN	RESPO	PATTE MCNAU	
AC066039B1069WHITC	RESPO	PAIRA ACOUS	
AC066039B1077TEAS	INTER TIME	INTEN TRADE	L, G,J,
AC066039B1086BREMN	NOT CLASSIFIABLE		
AC066039C1090	NOT CLASSIFIABLE		
AC066039D1102	LINE		M, M,
AC066039G1111GELLE	ULTRA	ATTEN SOLID	
AC066039G1120YCUSS	NOT CLASSIFIABLE	HIGH TEMPE	
AC066039E1125	NOT CLASSIFIABLE		
AC066039H1133	GUIDE SCUND	PROPA SHALL	H,
AC066039H1139PIERC	RELAT	CIRCU FLANE	B, H,J,K, G,H, B,F,L, B, G,
AC066039H1142WILLI	VELOC	POTEN DECRE	
AC066039I1145HERRM	CYLLIN	THICK	I,C, D,J, I,
AC066039I1154SCHAR	VIBRA		
AC066039J1162SPITZ	SOUND	REFLE PRESS	
AC066039J1170LENHA	SCUND	PROPA SHALL	
AC066039J1174MILNE	SHORE	FAST SEA	
AC066039B1183PCWER	LAIA	LOGGI AUDIO	
AC066039B1184KNIGH	THRES	DETER MANUA	
AC066039B1185CCLES	THRES	PURE	
AC066039B1187DELAN	THRES	SCUND	
AC066039F1188WILLI	MEASU	REACT TIME	
AC066039F1189FRICK	SINGH	EFFEC	
AC066039G1190	NOT CLASSIFIABLE		
AC066039J1191SCOTT	SEA	ICE	
AC066039L1193GOODE	SIGNA	RECEI	
AC066039L1193GOODE	NOISE	FIELD	
AC066040B	NOT CLASSIFIABLE		
AC066040B	CCCHL	MICRO	
AC066040B	CCMFO	CCCHL	
AC066040B	SUBHA	SHIFT	
AC066040B	DIRKS	LEVEL	
AC066040B	EGAN	LATER	
AC066040B	SCHE	FACTO	
AC066040B	ELFNE	SIGNA	
AC066040B	32MILNE	HARRI	
AC066040B	43HARRI	MASKF	
AC066040B	47FCOLE	HOCF	
AC066040B	54ETER	KLATI	
AC066040B	62MURLA	LAURE	
AC066040B	71HELLM	SCHAR	
AC066040B	79SCHRO	RESID	
AC066040B	INPUT	CHARA	
AC066040B	PITCH	MEMOR	
AC066040B	TOLER	REEXA	
AC066040B	RABIN	FURTH	
AC066040B	MODEL	CCCHL	
AC066040B	REFSOL	APPLI	
AC066040B	FOSSI	EXPLA	
AC066040B	LINE	SENSE ORGAN	
AC066040B	CLINI	PHYSI	
AC066040B	UNMAS	EC	
AC066040B	IMPAT	EARS	
AC066040B	EXPLA		

	CORRECT CATEGORY	ASSIGNED CATEGORIES	
		1ST RANK	2ND RANK
AC066040C 82REDWC LLOYD FINIT DIFFE METHC INVES VIBRA SOLID EVALU EQUIV C1ACU CHARA PIEZO RESON III CCNTO MODES RECTA PLATE Y-CUT QUART	C	I,	G, H,
AC066040C 86			
AC066040D 98			
AC066040E 108			
AC066040F0121ALLEN WESTE DIGIT COMPR TIME CORRE MATCH FILTRE ACTIV SONAR	F	F,	A,
AC066040F 123PAUL HOUSE STEVE ACCUS DESCR SYLLA NUCLE INTER TERMS DYNAM MODEL	F	B,	J,
ARTIC			
AC066040G 133			
AC066040G 138GFILLE ULTIFA PULSE PROPA THROU	G	G,	M,
AC066040G 148			
AC066040H 160			
AC066040H 163LORD CHANG VELOC ELAST PULSE	H	H,	
AC066040H 170			
AC066040H 176			
AC066040I 179RAND DIMAG AXISY VIBRA PROLA SPHER SHELL	I	I,	H,
AC066040I 187BIASI DITAR COMPO LOSS FACTO SELEC LAMIN BEAMS	I	I,	
AC066040I 195SPOLL GENER MATRI METHO DESIG ANALY VIBRA ISOLA SYSTE	I	I,	G, H,
AC066040J 205ARASE ARASE CORRE AMBLE SEA NOISE	J	J,	
AC066040J 211RUDNI BARMA KAGIW WANG FINCH STUDI THRES CAVIT NOISE LIQUI HELIO	J	J,	B,
AC066040K 219GULD HEAT TRANS ACROS SOLID LIQUI INTER PRESE ACOUS STREA	K	K,	G,
AC066040K 226			
AC066040K 229VORTH AIR BLAST SUPPR FUNCT EXELC CHARG BURIA DEPTH	K	K,	
AC066040K 240			
AC066040B 244BROAD TWO STAFF THRES MODEL RATING SCALE EXPR	B	B,	J,
AC066040B 245DECLAN DEATH HENDE EXTEN EXAMI	B	B,	
AC066040C 246			
AC066040D 247			
AC066040D 249PIERC ATIAI CCNSO ARBIT SCALE			
AC066040H 249YAMAD FUJII ACCUS RESPO RECTA	D	F,	D,
AC066040H 251GREEN BAFEL FISTO RALIA EXFAN	H	H,	J,
AC066040J 252			B,
AC066040J 254			
AC066040J 255HURDL FINE STRUC ACOUS FIELD	J	J,	
AC066040J 257			
AC066040J 257			
AC066040F 307REDDY SEGME SPEC SOUND	F	F,	B,
AC066040G 313BUCAR DARLY CARGM HUNTE ULTRA HYPER STUDI VIERA RELAX BENZE	G	G,	I,
AC066040G 317LCRBI TEMKI ATTEN DISPE SCUND PARTI RELAX PROCE	G	G,	
AC066040H 325EUPKE LONG WAVEL SCATT HARD SPHER	H	H,	
AC066040H 331			
AC066040H 342HAYEK VIBRA SPHER SHELL ACOUS MEDIO	H	I,	H,
AC066040H 349MANGU ACCUS RADIA WOEBL FISTO	H	H,	J,
AC066040H 354			
AC066040I 367AMBAR METHO CALCU FREQU PARTI	I	I,	
AC066040I 372CALLA BAKSH FLEXU VIERA C1ACU	I	I,	
AC066040I 376HILL TORSI WAVE PROPA RIGID SPHER SEMIE ELAST HALF SPACE	I	J,	C,
AC066040I 380FAKHO CCUFL VIERA SYSTE ANALY USING DUAL FORMA	I	I,	I,
AC066040I 385MURRA MCK FREE VIERA SIENE BAR NCNUN CHARA	I	I,	G, H,
AC066040I 390			B,

CORRECT CATEGORY		ASSIGNED CATEGORIES	
		1ST RANK	2ND RANK
ACO66040I	393SCOTT MIKLC CROSS SECTI		
ACO66040J	399BROWN REVER UNDER ARCTI SEA ICE	I,	H,
ACO66040J	405NARSH BRCWN SCATT LAYER WESTE NORTH ATLAN	J,	G,A,
ACO66040J	412 BEVER DEEP NOT CLASSIFIABLE	J,	A,
ACO66040J	417GOODM FREY ACOUS SCATT FLUID SPHER	J,	
ACO66040K	421RYAN WILLI SMITH GOULD MEASU STRUC HARMO SELF GENE ACOUS BEAM	K,	
ACO66040A	428WEST SESSL AVAL EVALU ACCUS PROPE ENCLO MEANS DIGIT COMPU	A,	C,
ACO66040A	434WEST SESSL AVAL SCHEO ACCUS MEASU PHILH HAIL NEW YORK	A,	C,
ACO66040B	441GLUCK BLACK MATUZ BAUER NCISE LOCAL UNILA ATTEN	J,	B,
ACO66040B	445LASKY CANPU INSTR METHO IMPRO INTEN DISCR DATA	B,	L,
ACO66040B	447 NOT CLASSIFIABLE		
ACO66040B	456WATSO FLANA BINAU UNMAS COMPL SIGNA	B,	L,
ACO66040B	469GEISL FRISH PERIP ORIGI AULIT RESPO RECOR EIGHT NERVE BULLE	B,	
ACO66040B	473 NOT CLASSIFIABLE		
ACO66040D	478WARD TEMPO THRES SHIFT MALES FEMAL	B,	D,
ACO66040C	486 NOT CLASSIFIABLE		
ACO66040E	496YCUNG ENERG SPECT DENSI SONIC BOCM	E,	D,
ACO66040H	498CADDE MIKI MFASU ROCM TEMPE MICRO ULTRA ATTEN Z-CUT QUART BRILL	G,	M,
ACO66040J	499ARASE MAPPI SPACE TIME CORRE AMBIE SEA NOISE		
ACO66040J	500MELLE IMIUL PROPA UNDER SCUND CHANN	J,	B,F,L,
ACO66040N	545MCCUE AURAL PULSE CCMER BAI5 HUEAN	J,	M,
ACO66040A	549 NOT CLASSIFIABLE		
ACO66040B	552BOER INTEN DISCR FLUCT SIGNA	B,	L,
ACO66040B	561LINNE CALIC EVEN ORDER SUEHA PERIP AUDIT SYSTE	B,	
ACO66040B	565LAU DESCR ANALY DOPPL DISCR FUNCT VARIA DIMEN SONAR ECHO	B,	L,
ACO66040B	570SMALL ACCLE TIME SEPAR PITCH ASSCC NOISE PULSE	B,	P,
ACO66040B	583GANNO MOSCC EFFEC CALCI SCUND EVCKE COCHL POTEN GUINE PIG	B,	H,
ACO66040B	591CARHA OLSEN INTEG ACCUS POWER THRES NCENA HEARE	B,	C,
ACO66040B	600GUTTM SCNDH WIDTH SPICT EFFEC BINAU RELEA MASKI	B,	F,J,
ACO66040C	607WALKE LORD ATHER BASIS IESIG CIRCU EAPRH SUITA MAP DETER	B,	C,J,
ACC66040F	614HOFFE SETHY SCHWE CHRIS NEW CORRE VOCOD	F,	
ACO66040F	621GCLDE IMFO NATUR INTEL EFFIU OXYGE SPEED USING VOCOD TECHN	F,	B,
ACO66040F	625MACLE ANALY SPEED HELIU OXYGE MIXTU UNDER PRESS	F,	J,
ACO66040F	628 NOT CLASSIFIABLE		
ACO66040F	635SINGH CROSS STUDY PERCE CCNPU PLCSI PHONE TWO CONDI DISTO	F,	
ACO66040F	657LAERI VOGEL WEISS IMPIE PITCH EXTRA DOUBL SPECT ANALY TYPE	F,	F,
ACO66040G	663SCHUL OCCNN BRILL SCATT TFERM RELAX BENZE	B,	
ACO66040H	667ENELE SCATT ARRAY CYLIN FURAM MOVIN ELEM	G,	H,J,K,
ACO66040I	671DIMOF ELECT VIERA STAND CERAM MOVIN ELEM	I,	J,L,
ACO66040I	677WEING PREST MAIAK JACCB VIEHA RECTA SANDW HONEY PLATE MASS ATTAC	I,	G,H,
ACO66040I	684KESSE RESON EXCIT ELAST CCNNE DCUBL BEAM SYSTE CYCLI MOVIN LCAD	I,	
ACO66040J	688GGCID MEASU VCLUM SCATT DEEP SCATT LAYER	J,	H,I,K,
ACO66040J	70697MARTI SEA SURFA ROUGH ACCUS REVER OPERA MODEL	J,	B,
ACO66040J	711 NOT CLASSIFIABLE		
ACO66040K	721RIERNE RESPO FLEXI PANFL TURBU FLOW RUNNI WAVE VERSU MODAL DENSI	K,	
ACO66040	727PATCH STAND REFER PRESS UNDER SOUND		J,
ACO66040	728YCUNG REFER PRESS SOUND PRESS LEVEL		F,
ACO66040G	728 NOT CLASSIFIABLE		

	CORRECT CATEGORY	ASSIGNED CATEGORIES	
		1ST RANK	2ND RANK
AC066040H 729TRUES EXIST LONGI WAVES			
AC066040H 730KCLD EXIST LONGI WAVES ANISO MEDIA	H	H	
AC066040K 731			
AC066040I 773EISNE INVER DESIG FLEXU VIBRA	H	H	
AC066040I 776IANG VIERA THICK CAELE	I	I	C, G, H,
AC066040I 784SACKM SHAH MCNIV AXIAL	I	I	
AC066040I 793IANG RESEO VISCO CYLIN	I	I	H, J, K,
AC066040I 801WILKI NATUR FEEQU CISCSE SPEER SANDW SHELL	I	I	H,
AC066040I 807YEH SLOSH LIQDI CONNE CYLIN TANKS CWING	I	I	G,
AC066040J 813BHAM ANAIC CCMPU PROGR STUDY UNDER SCUND RAYS U-TUB FREE OSCIL	J	J	B, F, L,
AC066040K 821PCHEL WHITE TRANS RANDO SCUND VIBRA THROU RECTA DOUBL WALL	K	I	G, H,
AC066040B 833GREEN SIGNA DETEC ANALY EQUAL CANCE MODEL	B	B	J,
AC066040B 839MCCCM HGDGE RELIA TTS IMPUL NOISE EXPCS	B	B	E,
AC066040G 847			
AC066040G 852BATEM MASON RELAT BETWE THIRD ORDER ELAST MODUL THERM ATTEN ULTRA	G	G	I,
AC066040G 863PAPAD WAVES NONCO METAL CRYST			
AC066040H 877SUN ULTRA LIEFR LCSS PHASE CHANG ANISO MATER	G	G	M,
AC066040H 883TWERS BURKE SCATT REFLE EILIP SPHER SURFA TIME DEPEN RADIU	H	H, H,	
AC066040H 896			
AC066040H 906			
AC066040B 911MCCOM HODGE ACCUS HAZAR CHILD TOYS	B	B	I,
AC066040B 914BRYAN TEMPE OBJEC AUDIO	B	B	
AC066040H 914			
AC066040I 915RAYNO VIERA VISCO ROD MEDIU PRODU VISCO DAMPL	I	I	G, H,
AC066040J 915WILSO FORMA DIFFE PATIE TRANS ODD NUMBE ELEME	J	I, G,	J, L, M,
AC066040M 916JANES ACKER CERIN EFFEC HEAT ULTRA VX-2 CARCI BONES RABBI PRELI	M		
AC066040C 949			
AC066040F 950			
AC066040F 955			
AC066040F 966			
AC066040E 979OHMAN PERCE SEGME VCCV UTTER	F	F, G,	M,
AC066040G 989FITCH NEW METHO MEASU ULTRA	G	G	
AC066040G 998IANG GROUP VELOC LISPE LUE PULSE REFLE FREQU DEPEN BOUND IMPED	G	G	M,
AC066040G 1002STEPH STERN SMITH THIRD CEDER ELAST MODUL POLYC METAL ULTRA VELOC	G	G	
AC066040G 1009			
AC066040G 1016DOBBT TEMKI MEASU ATTEN DISPE SCUND AEROS	G	G	J, G, H,
AC066040H 1025JUNGE ENERG EXCHA BETWE INCOM NEAR ACOUS FAR FIELD TRANS SCURC	H	I,	
AC066040H 1031LAURA DIREC CHARA VIERA CIRCU FLATE MEMBR	H	I,	
AC066040H 1034			
AC066040I 1039ROYST CLAYT IN FLANE INEXT VIERA CIRCU RING	I	I,	G, H,
AC066040I 1045EERNI COUEL MODE APPRO ELAST VIBRA ANALY	I	I,	G, H,
AC066040I 1051HOLLA NUNER STUDI ELAST DISK CONTO MODES LACKI AXIAL SYMME	I	I,	
AC066040I 1058KAPUR VIERA TIMOS BEAM USING FINIT ELEME APPRO	I	I,	J, L,
AC066040J 1064			
AC066040I 1073SHAH SACKM MCNIV AXIAL SYMME WAVES HOLIC ELAST RODS PART II	I	I,	
AC066040I 1077PAUL VIERA CIRCU CYLIN SHELL FIEZO SILVE IODIE CRYST	I	I,	H,

	CORRECT CATEGORY	ASSIGNED CATEGORIES									
		1ST RANK	2ND RANK								
AC066040I1081WANG	GRAPH METHO	OBTAI	NATUR	FREQU	THRFE	DEGRE	FREED	SYSTE	I		
AC066040J1083MANGU	ACCUS	RADIA	FISTO	IAYLR	MEDIU	APPLI	LAYER	CONTA	BUBBL	J	H,
AC066040J1094CHON	NUTTA	SIGNA	WAVEF	CISTC	CAUSE	REFLE	LOSSY	LAYER	BOTTO	J	B,
AC066040J1108URICK	CORRE	PROPE	AMBIE	NOISE	BERMU					J	
AC066040J1112						NOT CLASSIFIABLE				K	H,J,K,
AC066040K1124WHITE	SCUND	THANS	THRCU	FINIT	CLCSE	CYLIN	SHELL			B	
AC066040B1131CAPRA	VOCAL	RESPO	BULLE	NATUR	SYNTH	MATIN	CALLS			B	F,M,
AC066040B1140CLACK	EFFEC	SIGNA	LURAT	AUDIT	SENSI	HUMAN	MONKE			B	J,
AC066040B1147CRANE	MECHA	IMPAC	MODEL	AUDIT	EXCIT	FATIG				B	D,J,
AC066040B1160JCHNS	DALLO	INFLU	RISE	FAIL	TIME	UPON	SHORT	THRES		B	C,
AC066040B1164VENDR	THLJS	FIJKM	WEER	LAW	PCWER	LAW	INTER	NOISE		B	D,J,
AC066040B1174CAMPB	HOTLO	STUCK	DECIS	RULES	THRES	DETER				B	F,
AC066040B1180SWIGA	PITCH	PERIO	INIER	TCNE						B	I,
AC066040B1186CLACK	BEHAV	PINNA	REFLE	THRES	RAT	AUDIO	TTS	COVAR		B	
AC066040B1187TRAUT	GAREA	TIME	FHEQU	ANALY	HEARI	PROCE				B	I,
AC066040B1189IYBAR	INTER	RCNE	CCNEU	THRES	AUDIO					B	
AC066040B1190SHEPO	RIACH	SUFPL	OBSER	CENTR	FACIO	AUDIT	FATIG			B	L,
AC066040B1192SCHAR	CCMMF	MPSKI	DISCH							B	J,
AC066040H1193FUJII	YAMAD	ACOUS	RESPO	CIRCU	RECEI	CIRCU	SCURC	DIFFE	RADIU	H	
AC066040J1195ARASE	CCMME	ACOUS	RESPO	RECTA	RECEI	RECTA	SOURC			J	H,
AC066040J1195YARNA	DANN	CLARK	RECEN	RESUL	STRAI	FLORI	UNDER	SOUND	PROPA	J	B,P,L,
AC066040J1197WAREN	SUCHM	LASDO	NONLI	RECGR	APPLI	LINEA	ARRAY	DESIG		J	G,
AC066040J1200MEILE	UNDER	ACOUS	SCATT	ARCIL	ICE					J	B,
AC066040J1202WATSO	MODEL	SOUND	CHANN	SINUS	RAYS					J	M,
AC066040M1202LUNN	NACIE	ULTRA	IRRAD	ENZYM	SCLUT					G	
AC066040 1203ERRAT	TABLE	CCNTE									P,
ACC66040 1287EXPER	BIELI	REFER								J	
AC066040J1288BROCK	BAXTE	SCME									
AC066040J1300ECRBE	DIFFR	EFFEC								J	H,
AC066040J1305CHIN	NEAR	FIELD								J	I,
AC066040J1317SCHUY	RALIA	RESIS									
AC066040J1337DENHA	ARRAY	RIGID	FLANE							J	
AC066040J1331	GOLD	TIME									
AC066040J1345SABIN	KIBBL	EXPER								J	H,
AC066040J1354WHITE	ACCUS	IMPED								J	
AC066040M1363SICHE	TRANS	BOUND								K	J,
AC066040B1371LABEN	SCHNI	NYBOR	WIERC	WILSO	DEFECR	MOTIC	PRODU	ISOLA	LIVIN	G	M,
AC066040B1381EALLO	ICCCL	ULTRA	VIERA								D,
AC066040B1392EURLA	KYLIN	COHEN	TEMPO	THRES	SHLFT	HEARI	EXPOS	COMBI	IMPAC	B	
AC066040B1398FUGSL	STATE	NOISE	CONDI							B	J,
AC066040B1405JOHNS	GENER	ODD	FRACT	SURHA						B	
AC066040B1414MCFAD	APPLI	FC	MODEL	INTER	JNDS					B	
AC066040B1420SEIRA	JOHNS	JCHNS	MEMER	RESIS	ENDCL	WALLS	FIRST	TURN	GUINE	B	
AC066040B1405JOHNS	CCCHL										
AC066040B1414MCFAD	JCHNS	ORIGI	SUMMA	FCTEN						B	H,
AC066040B1420SEIRA	MASKI	LEVEL	DIFFE	CCNTI	BURST	MASKI	NOISE			B	P,J,
AC066040B1420SEIRA	FREI	FISCH	FELLM	RUBIN	MEASU	STAPE	FOOTP	DISPL	DURIN	B	B,
AC066040B1420SEIRA	SCUND	THECU	MIDEL	EAR							

B30024